Mathematische Methoden der Unsicherheitsquantifizierung

Oliver Ernst

Professur Numerische Mathematik

Sommersemester 2014

TECHNISCHE UNIVERSITÄT
CHEMNITZ

Organisatorisches

- Vorlesungswebseite:
  `www.tu-chemnitz.de/mathematik/numa/lehre/uq-2014`
- Vorlesung: Prof. Oliver Ernst,
  `oliver.ernst@mathematik.tu-chemnitz.de`
  Mo 13:45 & Mi 15:30
- Übung: Dipl.-Math. Björn Sprungk,
  `bjoern.sprungk@mathematik.tu-chemnitz.de`
  Mo 15:30
- Prüfung: 30 Minuten mündlich, Termin nach Vereinbarung.
- Modul FN3: 8 LP, 240 AS.

Wesentliche Inhalte

Folgende Themen werden bahandelt:

- Zuvallsvariable mit Werten in abstrakten Räumen
- Darstellung von Zufallsfeldern
- Monte Carlo Methoden
- Kollokation bzw. hochdimensionale Quadratur und Interpolation
- Dünne Gitter
- Polynomielle Chaosentwicklungen
- Stochastische Galerkin Diskretisierung

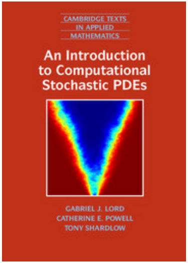
Nicht behandelt werden

- Stochastische Differentialgleichungen (völlig andere Methoden, siehe Stochastik-LV)
- Inverse Probleme (vielleicht im nächsten Durchlauf)

Literatur

Wir folgen teilweise dem (bald erscheinenden) Buch

An Introduction to Computational Stochastic PDEs
von Lord, Powell und Shardlow
Cambridge University Press, 2014.



CAMBRIDGE TEXTS IN APPLIED MATHEMATICS

An Introduction to Computational Stochastic PDEs

GABRIEL J. LORD
CATHERINE E. POWELL
TONY SHARDLOW

Weitere Bücher zum Thema:

- D. Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach.* Princeton University Press, Princeton, NJ, 2010.
- O. P. Le Maître and O. M. Knio. *Spectral Methods for Uncertainty Quantification.* Springer-Verlag, Dordrecht Heildelberg, 2010.
- R. C. Smith. *Uncertainty Quantification: Theory, Implementation and Applications.* Computational Science and Engineering. SIAM, 2014.
- R. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach.* Springer-Verlag, New York, 1991.

Siehe auch die (laufend ergänzte) Literaturliste auf der Webseite.

Hilfreiches Vorwissen

- Grundlagen Numerik
- Grundlagen Stochastik (wird aufgefrischt)
- Grundlagen Funktionalanalysis
- Theorie und Numerik elliptischer Randwertaufgaben (FE)

Contents

Contents

1. Introduction

2. Monte Carlo Methods

3. Random Fields

4. Stochastic Collocation

Contents

1 Introduction

1.1 What is Uncertainty Quantification?

1.2 A Case Study: Radioactive Waste Disposal

2 Monte Carlo Methods

3 Random Fields

4 Stochastic Collocation

What is Uncertainty Quantification? (UQ)

- What is uncertainty quantification (UQ) about?

- What is uncertainty?
- How can uncertainty be described?
- How can the effects of uncertainty be treated and quantified?
- Methods for solving the resulting mathematical problems.

What is Uncertainty Quantification? (UQ)

What is 'uncertain'?

uncertainty: *Not able to be relied on; not known or definite.*

Oxford Collegiate Dictionary

uncertainty: *not exactly known or decided; not definite or fixed*

Merriam Webster Online Dictionary

What is Uncertainty Quantification? (UQ)

Auf Deutsch?

unsicher: *gefahrvoll, gefährlich, keine Sicherheit bietend*
gefährdet, bedroht
das Risiko eines Misserfolges in sich bergend, keine [ausreichenden] Garantien bietend; nicht verlässlich; zweifelhaft unzuverlässig
einer bestimmten Situation nicht gewachsen, eine bestimmte Fähigkeit nicht vollkommen, nicht souverän beherrschend nicht selbstsicher
(etwas Bestimmtes) nicht genau wissend
nicht feststehend; ungewiss

Duden Online

What is Uncertainty Quantification? (UQ)

Auf Deutsch?

unsicher: *gefahrvoll, gefährlich, keine Sicherheit bietend*
gefährdet, bedroht
das Risiko eines Misserfolges in sich bergend, keine [ausreichenden] Garantien bietend; nicht verlässlich; zweifelhaft unzuverlässig
einer bestimmten Situation nicht gewachsen, eine bestimmte Fähigkeit nicht vollkommen, nicht souverän beherrschend nicht selbstsicher
(etwas Bestimmtes) nicht genau wissend
nicht feststehend; ungewiss

Duden Online

ungewiss: *fraglich, nicht feststehend; offen*
unentschieden, noch keine Klarheit gewonnen habend
(gehoben) so [beschaffen], dass nichts Deutliches zu erkennen, wahrzunehmen ist; unbestimmbar

Duden Online

What is Uncertainty Quantification? (UQ)

A poetic description

There are known knowns;
there are things we know we know.

We also know there are known unknowns;
that is to say, we know there are some things we do not know.

But there are also unknown unknowns – the ones we don't know we don't know.

U. S. Secretary of Defense, Donald Rumsfeld
DoD News Briefing; Feb. 12, 2002

What is Uncertainty Quantification? (UQ)

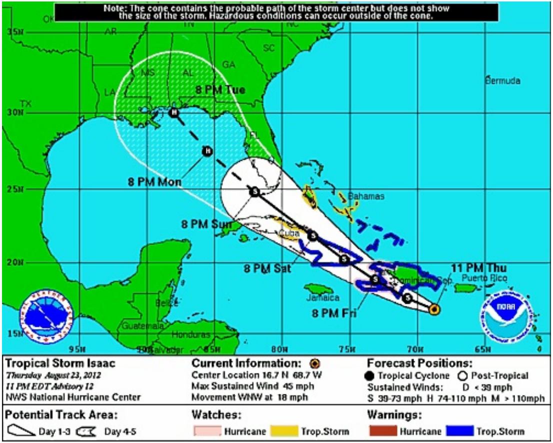
Uncertainty in Modern Life

(Increasingly?) many aspects of modern life involve uncertainty.

- **Social systems:** military, finance, insurance industry, elections
- **Environmental systems:** weather, climate, seismics, subsurface geophysics
- **Engineering systems:** automobiles, aircraft, bridges, structures
- **Biological systems:** health and medicine, pharmaceuticals, gene expression, cancer research
- **Physical systems:** quantum physics, radioactive decay

What is Uncertainty Quantification? (UQ)

Uncertainty in Modern Life

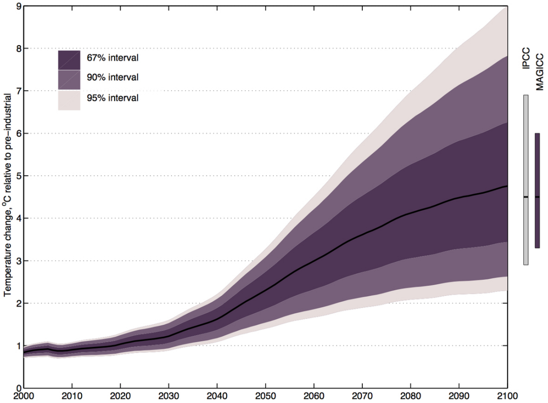


Source: National Hurricane Center, USA

Predicted storm path with uncertainty cones.

What is Uncertainty Quantification? (UQ)

Uncertainty in Modern Life

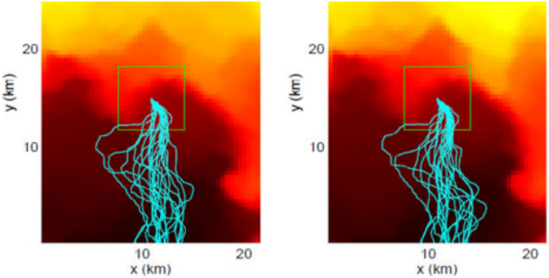


Source: Brodman & Karoly, 2013

Global-mean temperature change for a business-as-usual emission scenario, relative to pre-industrial. Black line: median, shaded regions 67% (dark), 90% (medium) and 95% (light) confidence intervals.

What is Uncertainty Quantification? (UQ)

Uncertainty in Modern Life



Source: K. A. Cliffe, 2012

Sample paths of groundwater-borne contaminant particles emanating from an underground radioactive waste disposal site.

What is Uncertainty Quantification? (UQ)

Examples

Radioactive decay

- Radium-226: half-life of 1602 years
- Decays into Radon gas (Radon-222) by emitting alpha particles.
- Over a period of 1602 years, half the radium atoms in a given sample will decay.
- But we cannot say which half!

This kind of uncertainty seems to be ‘built in’ to the physical world.

What is Uncertainty Quantification? (UQ)

Examples

Rolling a die (or several dice)

- Cube, 6 faces, numbered 1–6
- One or more thrown onto a table.
- For “fair dice”, expect to see the numbers 1–6 appear equally often, provided the dice are thrown sufficiently many times.

How does this differ from radioactive decay?

Is this uncertainty also 'built-in' to the physical world, or is it just that we don’t know how to calculate what will happen when the dice are thrown?

What is Uncertainty Quantification? (UQ)

Examples

Screening/testing for disease

- Incidence of disease among general population: 0.01 %
- Test has true positive rate (sensitivity) of 99.9 %.
- Same test has true negative rate (specificity) of 99.99 %.
- What is the chance that someone who tests positive actually has the disease?

Screening/testing for disease

- Incidence of disease among general population: 0.01 %
- Test has true positive rate (sensitivity) of 99.9 %.
- Same test has true negative rate (specificity) of 99.99 %.
- What is the chance that someone who tests positive actually has the disease?

Answer (relative probabilities, conditional probabilities, Bayes' formula)

$$\begin{aligned}\mathbf{P}(\text{desease}|\text{pos}) &= \frac{\mathbf{P}(\text{pos}|\text{disease}) \cdot \mathbf{P}(\text{disease})}{\mathbf{P}(\text{pos}|\text{disease}) \cdot \mathbf{P}(\text{disease}) + \mathbf{P}(\text{pos}|\text{no disease}) \cdot \mathbf{P}(\text{no disease})} \\ &= \frac{0.999 \cdot 0.0001}{0.999 \cdot 0.0001 + (1-0.9999) \cdot (1-0.0001)} \\ &\approx 0.4998\end{aligned}$$

What is Uncertainty Quantification? (UQ)

Examples

Alternative answer (natural frequencies)

- Think of random sample 10,000 people.
- Of these, on average 1 will have the disease, 9,999 will not.
- The person who has the disease will almost certainly test positive.
- of the 9,999 healthy people, on average one will test (falsely) positive.
- Thus, roughly one out of every two positive patients actually has the disease.

What is Uncertainty Quantification? (UQ)

Alternative answer (natural frequencies)

- Think of random sample 10,000 people.
- Of these, on average 1 will have the disease, 9,999 will not.
- The person who has the disease will almost certainly test positive.
- of the 9,999 healthy people, on average one will test (falsely) positive.
- Thus, roughly one out of every two positive patients actually has the disease.

In [Gigerenzer, 1996] medical practitioners were given the following information regarding mammography screenings for breast cancer:

$$\text{incidence: } 1\ \%; \quad \text{sensitivity: } 80\ \%; \quad \text{specificity: } 90\ \%.$$

When asked to quantify the probability of the patiant actually having breast cancer given a positive screening result (7.5%), 95 out of 100 physicians estimated this probability to lie above 75%.
See also [Gigerenzer et al., 1998] for similar observations in AIDS counseling.

What is Uncertainty Quantification? (UQ)

Examples

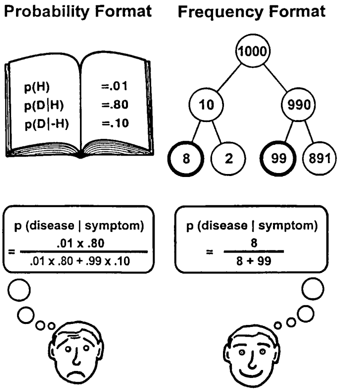

Probability Format
Frequency Format
p(H) =.01
p(D|H) =.80
p(D|-H) =.10
1000
10
990
8
2
99
891
p (disease | symptom) = .01 x .80 / .01 x .80 + .99 x .10
p (disease | symptom) = 8 / 8 + 99

FIGURE 1. Bayesian computations are simpler when information is represented in a frequency format (right) than when it is represented in a probability format (left). p(H) = prior probability of hypothesis. H (breast cancer), p(D|H) = probability of data D (positive test) given H, and p(D|−H) = probability of D given −H (no breast cancer).

Sometimes the description of uncertainty is crucial for its transparent communication.

Source: Gigerenzer, 1996

What is Uncertainty Quantification? (UQ)

Examples

Modeling biological systems

- From one view, biology is just very complicated physics and chemistry.
- But even the simplest biological systems are far too complicated to be understood from basic principles at the moment.
- Models are constructed that attempt to capture the essential features of what is happening, but often there are competing models and they may all fail in some way or other to predict the observed phenomena.
- In short, we don't really know what the model is!

How does this situation differ from the previous two?

Climate change

> *The weight of evidence makes it clear that climate change is a real and present danger. The Exeter conference was told that whatever policies are adopted from this point on, the Earth's temperature will rise by 0.6F within the next 30 years. Yet those who think climate change just means Indian summers in Manchester should be told that the chances of the Gulf stream - the Atlantic thermohaline circulation that keeps Britain warm - shutting down are now thought to be greater than 50%.*
>
> *The Guardian, 2005*

> *Most of the observed increase in globally-averaged temperatures since the mid-20th century is very likely due to the observed increase in anthropogenic GHG concentrations. It is likely there has been significant anthropogenic warming over the past 50 years averaged over each continent (except Antarctica).*
>
> *IPCC Fourth Assessment*
> *Summary for Policymakers.*

What do these statements mean?

What is Uncertainty Quantification? (UQ)

Examples

Unknown unknowns

- Obviously can't give a current example.
- A good example ist the state of Physics at the end of the 19th century.

> *There is nothing new to be discovered in physics now. All that remains is more and more precise measurement.*
>
> *Lord Kelvin, 1900*

- Quantum mechanics and relativity theory were unknown unknowns.

It is easy to underestimate uncertainty.

What is Uncertainty Quantification? (UQ)

Political Implications

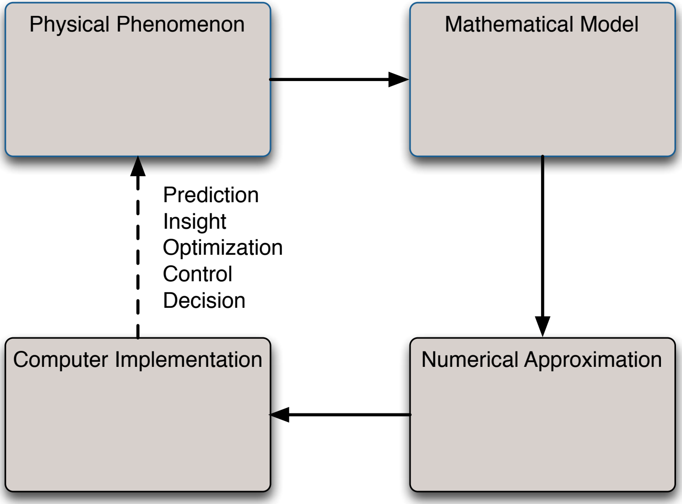
Questions:[1]

1. How do we account for all the uncertainties in the complex models and analyses that inform decision makers?
2. How can those uncertainties be communicated simply but quantitatively to decision makers?
3. How should decision makers use those uncertainties when combining scientific evidence with more socio-economic considerations?
4. How can decisions be communicated so that the proper acknowledgment of uncertainty is transpartent?

[1] posed on entry at the 2006 EPSRC Ideas Factory on the topic *Scientific Uncertainty and Decision Making for Regulatory and Risk Assessment Purposes*.

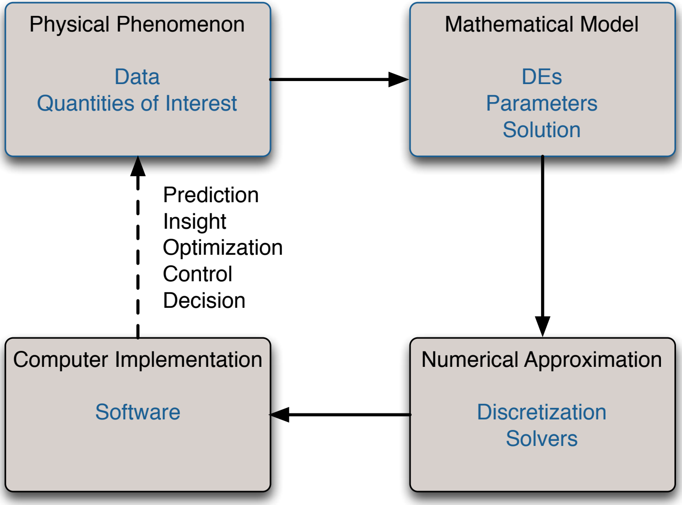
What is Uncertainty Quantification? (UQ)

UQ and the scientific computing paradigm



What is Uncertainty Quantification? (UQ)

UQ and the scientific computing paradigm



Physical Phenomenon

Data
Quantities of Interest

Mathematical Model

DEs
Parameters
Solution

Numerical Approximation

Discretization
Solvers

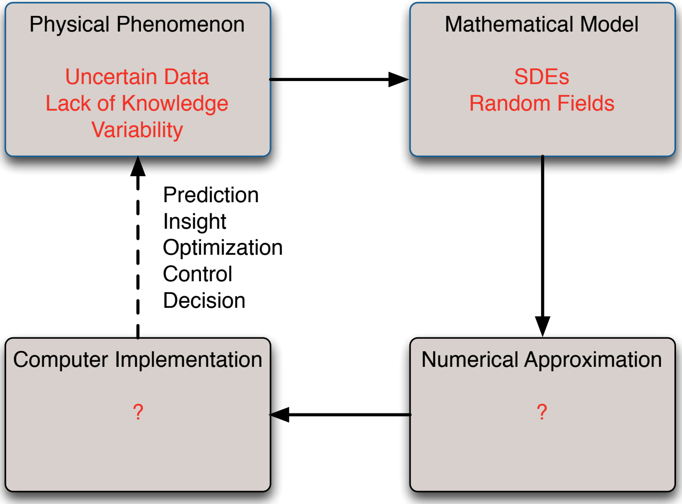
Computer Implementation

Software

Prediction
Insight
Optimization
Control
Decision

What is Uncertainty Quantification? (UQ)

UQ and the scientific computing paradigm


Physical Phenomenon

Uncertain Data
Lack of Knowledge
Variability

Mathematical Model

SDEs
Random Fields

Prediction
Insight
Optimization
Control
Decision

Computer Implementation

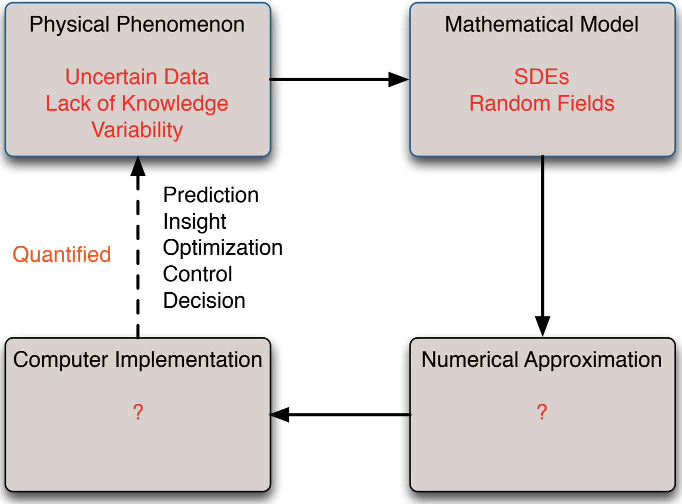
?

Numerical Approximation

?

What is Uncertainty Quantification? (UQ)

UQ and the scientific computing paradigm



Physical Phenomenon

Uncertain Data
Lack of Knowledge
Variability

Mathematical Model

SDEs
Random Fields

Numerical Approximation

?

Computer Implementation

?

Quantified

Prediction
Insight
Optimization
Control
Decision

What is Uncertainty Quantification? (UQ)

Validation and Verification (V & V)

What confidence can be assigned to a computer prediction of complex phenomena?

Validation: The determination of whether a mathematical model adequately represents the pysical or engineering phenomenon under study.
"Are we solving the right problem?"

Is this even possible? (cf. Carl Popper)

Verification: The determination of whether an algorithm and/or computer code correctly implements a given mathematical model.
"Are we solving the problem correctly?"

- code verification (software engineering)
- solution verification (a posteriori error estimation)

What is Uncertainty Quantification? (UQ)

Aleatory and Epistemic Uncertainty

Aleatoric Uncertainty: (variability) Uncertainty due to true intrinsic variability; cannot be reduced by additional experimentation, improvement of measuring devices etc.

Examples:

- rolling a die
- wind stress on a structure
- production variations

Epistemic Uncertainty: Uncertainty due to lack of knowledge/incomplete information.
Examples:

- turbulence modeling assumptions
- surrogate chemical kinetics
- the probability distribution a random quantity follows

Note: This distinction is not always meaningful or possible.

What is Uncertainty Quantification? (UQ)

Model Problem

The most popular model problem in the UQ community has become the second-order elliptic PDE with an uncertain coefficient function:

$$-\nabla\cdot(a\nabla u) = f \quad + \text{ domain } D \subset \mathbb{R}^d \quad + \text{ BC.}$$

Rather than the solution u (whatever that may be), typical problems in UQ require a functional Q of the solution, e.g. its value at a point in the computational domain. Such a functional is known as a quantity of interest (QoI).

Examples:

$$Q(u) = u(\boldsymbol{x}_0), \qquad Q(u) = \frac{1}{|D_0|}\int_{D_0} u(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Introduce associated output set $G = \{Q(u)\}$ for all possible solutions u.
Consider mapping $P : S \to G$ of all possible inputs to output set G.

In what way might uncertainty in the coefficient a be addressed?

What is Uncertainty Quantification? (UQ)

Worst case analysis

Introduce an ϵ-ball around a given function a_0 (in a suitable norm).

Examples:

$$\begin{aligned} S_\infty &:= \{a \in L^\infty(D) : \|a - a_0\|_{L^\infty(D)} \leqslant \epsilon\}, \\ S_1 &:= \{a \in W^{1,\infty}(D) : \|a - a_0\|_{W^{1,\infty}(D)} \leqslant \epsilon\}, \\ S_{\text{const}} &:= \{a : a \text{ is constant in } D, |a - a_0| \leqslant \epsilon\}. \end{aligned}$$

Worst case analysis: determine uncertainty interval

$$I = [\inf_{a \in S} Q(u(a)), \sup_{a \in S} Q(u(a))].$$

The uncertainty range of Q is then the length of I.

This is a generalization of interval analysis.

What is Uncertainty Quantification? (UQ)

Probabilistic model

Idea: Some values (functions) $a \in S$ are more likely than others.

Purely probabilistic approach:

- Introduce probability measure on S.
- (Measurable) mapping $P : S \to G$ induces probability measure on G. ("uncertainty propagation")
- Big issue: choice of distribution, too much subjective information?
- Some classical guidelines: Laplace's principle of insufficient reason, maximum entropy, etc.
- Choosing distribution based on data is point of departure for Bayesian inverse problem.

What is Uncertainty Quantification? (UQ)

Evidence theory

Generalizes probabilistic model (also called Dempster-Shafer theory)

- Finite or countable family $\mathfrak{F}$ of events.
- Set function $m : \mathfrak{F} \to [0,1]$ giving likelihood information for each event, satisfies
$$\sum_{A \in \mathfrak{F}} m(A) = 1, \qquad m(\emptyset) = 0,$$
but, unlike probability measures, need not satisfy $A \subset B \Rightarrow m(A) \leqslant m(B)$.
- Belief and plausability functions for admissible events C
$$\mathsf{bel}(C) = \sum_{A \in \mathfrak{F}, A \subset C} m(A), \qquad \mathsf{pl}(C) = \sum_{A \in \mathfrak{F}, A \cap C \neq \emptyset} m(A).$$
provide lower and upper bounds, respectively, on likelihood of event C.
- Likelihood function dependent on expert opinion.

What is Uncertainty Quantification? (UQ)

Fuzzy sets and possibility theory

Deterministic approach introduced by [Zadeh, 1965].

- Generalizes "$\in$" relation of classical set theory: for $C \subset S$, in place of exhaustive alternatives $x \in C$ and $x \notin C$, introduces membership function
$$\mu_C : S \to [0, 1]$$
expressing truth degree of statement $x \in C$.
- Important tool: α-cut of set C defined by
$$C^\alpha := \{x \in S : \mu_C(x) \geqslant \alpha\}$$
giving set characterization of uncertainty.
- Mapping P then again propagates fuzziness of input set S to output set G.

Contents

A Case Study: Radioactive Waste Disposal

- An area where UQ has played a central role in the past 25 years is the assessment of strategies and sites for the long-term storage of radioactive waste.
- Uncertainties arise from technological complexity as well as the long time scales to be considered.
- Many leading industrial countries (USA, UK, Germany) have scrapped previous plans for national long-term disposal sites and are re-evalutating their strategies.
- We consider a basic UQ problem which occurs in site assessment studies.

A Case Study: Radioactive Waste Disposal

Background

- Radioactive waste is produced in large part by power plants, in which the heat from controlled nuclear fission is used to produce electric power. (Other sources: medical, weapon production, non-nuclear industries)
- Exposure to high radiation levels seriously harmful to humans and animals; long-term exposure to low-level radiation can cause cancer and other long-term health problems.
- Classification
  - high-level waste (HLW): highly radioactive, produces heat, small quantities.
  - intermediate-level waste (ILW): still very radioactive, does not produce heat.
  - low-level waste (LLW): low radiactivity; packaging material, protective clothing, soil, concrete etc. which has been exposed to radioactivity.
- Quantities in storage (source: IAEA database, `http://newmdb.iaea.org`)
  - Germany: 120,000 m^3 (2007)
  - France: 90,000 m^3 (2007)
  - UK: 350,000 m^3 (2007)
  - USA: 540,000 m^3 (2008)

A Case Study: Radioactive Waste Disposal

Management Options

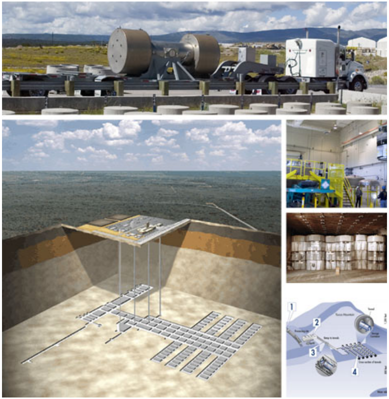
Since this problem has received serious consideration ($\approx$ 1970s), several options have been discussed

- Surface storage: current universal solution, not long-term, risky.
- Disposal at sea: banned by international treaty (London Convention)
- Disposal in space: too dangerous, prohibitive cost (but permanent solution).
- Transmutation: not yet proven technology, would mitigate but not solve the problem.
- Deep geological disposal
  - Favored by nearly all countries with a radioactive waste disposal program.
  - Storage in containers in tunnels, several hundred meters deep, in stable geological formations.
  - Issue: retrievable or not?
  - No human intervention required after final closure of repository.
  - Several barriers: chemical, physical, geological.
  - Substantial engineering challenge (containment must be assured for at least 10,000 years).
  - Main escape route for radionuclides: groundwater pathway.

A Case Study: Radioactive Waste Disposal

WIPP

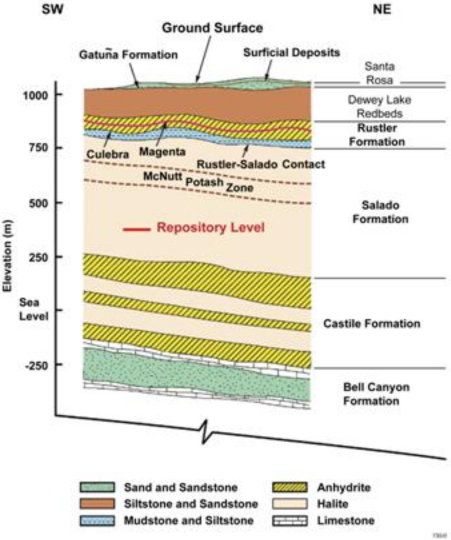
- US DOE repository for radioactive waste situated near Carlsbad, NM.
- Fully operational since 1999.
- Extensive site characterization and performance assessment since 1976, also in course of compliance certification and recertification by US EPA (every 5 years).
- Large amount of publicly available data.
- http://www.wipp.energy.gov



A Case Study: Radioactive Waste Disposal

WIPP geology

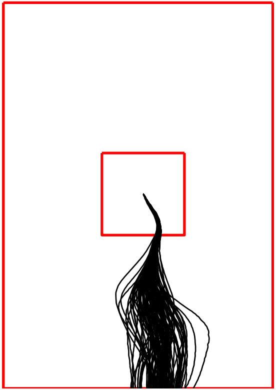
- Repository located at depth of 655 m within bedded evaporites, primarily halite (salt).
- The most transmissive rock in the region is the Culebra Dolomite.
- In the event of an accidental breach, Culebra would be the principal pathway for transport of radionuclides away from the repository.



A Case Study: Radioactive Waste Disposal

WIPP UQ scenario

- One scenario at WIPP is a release of radionuclides by means of a borehole drilled into the repository.
- Radionuclides are released into the Culebra Dolomite and then transported by groundwater.
- Travel time from release point in the repository to the boundary of the region is an important quantity.
- Flow is two-dimensional to a good approximation.

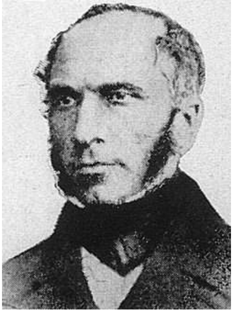


Darcy's law

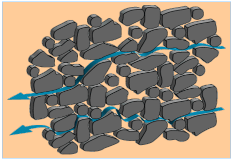
- The simplest mathematical model for flow through a porous medium (as in groundwater through an aquifer) is given by Darcy's Law

  $$\boldsymbol{q} = \frac{-\boldsymbol{k}}{\mu} \nabla p,$$

  in which $\boldsymbol{q}$ is the volumetric flux or Darcy velocity (discharge per unit area in [m/s]), $\boldsymbol{k}$ is the permeability tensor, a material parameter describing how easily water flows through the given medium, μ is the dynamic viscosity of the fluid and p is the hydraulic head of the fluid.



- The hydraulic conductivity tensor is defined as $K := \boldsymbol{k}\rho g/\mu$, where g is the acceleration due to gravity and ρ the fluid density.



- The actual pore velocity with which the fluid particles move through the pores is obtained as $\boldsymbol{u} = \boldsymbol{q}/\phi$, where $phi \in [0, 1]$ denotes the porosity of the medium.

A Case Study: Radioactive Waste Disposal

Groundwater Flow Model

Stationary Darcy flow	$\boldsymbol{q} = -K\nabla p$	$\boldsymbol{q}$: Darcy flux
		K : hydraulic conductivity
		p : hydraulic head
mass conservation	$\nabla\cdot\boldsymbol{u} = 0$	$\boldsymbol{u}$: pore velocity
	$\boldsymbol{q} = \phi\boldsymbol{u}$	ϕ : porosity
transmissivity	$T = Kb$	b : aquifer thickness
particle transport	$\dot{\boldsymbol{x}}(t) = -\dfrac{T(\boldsymbol{x})}{b\phi}\nabla p(\boldsymbol{x})$	$\boldsymbol{x}$: particle position
	$\boldsymbol{x}(0) = \boldsymbol{x}_0$	$\boldsymbol{x}_0$: release location

Quantity of interest: particle travel time to reach WIPP boundary (actually, its $\log_{10}$).

A Case Study: Radioactive Waste Disposal

PDE with Random Coefficient

Primal form of Darcy equations:

$$\nabla\cdot[T(\boldsymbol{x})\nabla p(\boldsymbol{x})] = 0, \quad \boldsymbol{x} \in D, \qquad p = p_0 \text{ along } \partial D.$$

Model T as a random field (RF) $T = T(\boldsymbol{x}, \omega)$, $\omega \in \Omega$, with respect to underlying probability space $(\Omega, \mathfrak{A}, \mathbf{P})$.

Modeling Assumptions: (standard in hydrogeology)

- T has finite mean and covariance

$$\begin{aligned} \overline{T}(\boldsymbol{x}) &= \mathbf{E}\left[T(\boldsymbol{x},\cdot)\right], & \boldsymbol{x} &\in D, \\ \mathbf{Cov}_T(\boldsymbol{x},\boldsymbol{y}) &= \mathbf{E}\left[\left(T(\boldsymbol{x},\cdot) - \overline{T}(\boldsymbol{x})\right)\left(T(\boldsymbol{y},\cdot) - \overline{T}(\boldsymbol{y})\right)\right], & \boldsymbol{x},\boldsymbol{y} &\in D. \end{aligned}$$

- T is lognormal, i.e., $Z(\boldsymbol{x},\omega) := \log T(\boldsymbol{x},\omega)$ is a Gaussian RF.
- $\mathbf{Cov}_Z$ is stationary and isotropic, i.e., $\mathbf{Cov}_Z(\boldsymbol{x},\boldsymbol{y}) = c(\|\boldsymbol{x}-\boldsymbol{y}\|_2)$, and of Matérn type.

A Case Study: Radioactive Waste Disposal

Matérn Family of Covariance Kernels

$$c(\boldsymbol{x}, \boldsymbol{y}) = c_{\boldsymbol{\theta}}(r) = \frac{\sigma^2}{2^{\nu-1}\,\Gamma(\nu)} \left(\frac{2\sqrt{\nu}\,r}{\rho}\right)^{\nu} K_{\nu}\left(\frac{2\sqrt{\nu}\,r}{\rho}\right), \quad r = \|\boldsymbol{x} - \boldsymbol{y}\|_2$$

K_ν : modified Bessel function of order ν

Parameters $\boldsymbol{\theta} = (\sigma^2, \rho, \nu)$

σ^2 : variance

ρ : correlation length

ν : smoothness parameter

A Case Study: Radioactive Waste Disposal

Matérn Family of Covariance Kernels

$$c(\boldsymbol{x},\boldsymbol{y}) = c_{\boldsymbol{\theta}}(r) = \frac{\sigma^2}{2^{\nu-1}\,\Gamma(\nu)}\left(\frac{2\sqrt{\nu}\,r}{\rho}\right)^{\nu} K_{\nu}\left(\frac{2\sqrt{\nu}\,r}{\rho}\right), \quad r = \|\boldsymbol{x}-\boldsymbol{y}\|_2$$

K_ν : modified Bessel function of order ν

Parameters $\boldsymbol{\theta} = (\sigma^2, \rho, \nu)$

σ^2 : variance

ρ : correlation length

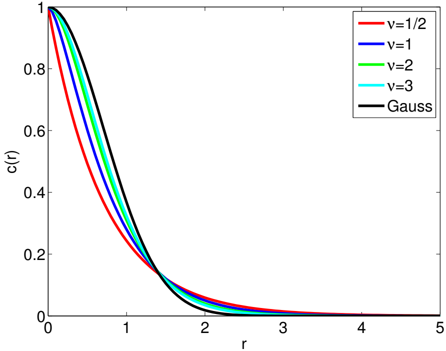
ν : smoothness parameter

Special cases:

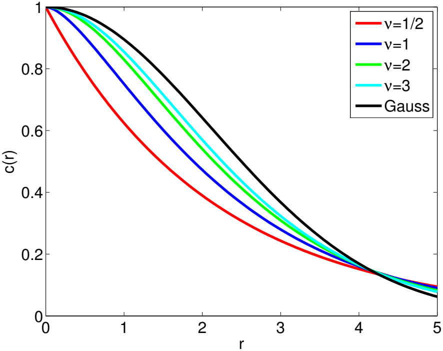
$\nu = \frac{1}{2}$:	$c(r) = \sigma^2 \exp(-\sqrt{2}r/\rho)$	exponential covariance
$\nu = 1$:	$c(r) = \sigma^2 \left(\frac{2r}{\rho}\right) K_1\left(\frac{2r}{\rho}\right)$	Bessel covariance
$\nu \to \infty$:	$c(r) = \sigma^2 \exp(-r^2/\rho^2)$	Gaussian covariance

A Case Study: Radioactive Waste Disposal

Matérn Covariance Functions



$\rho = 1$



$\rho = 3$

Smoothness: Realizations $Z(\cdot,\omega)$ are k times differentiable $\Leftrightarrow$ $\nu > k$.

A Case Study: Radioactive Waste Disposal

Karhunen-Loève expansion

Covariance function of RF $Z \in L^2_{\mathbf{P}}(\Omega; L^\infty(D))$

$$c(\boldsymbol{x}, \boldsymbol{y}) = \mathbf{Cov}_Z(\boldsymbol{x}, \boldsymbol{y}) := \mathbf{E}\left[\Big(Z(\boldsymbol{x}, \cdot) - \overline{Z}(\boldsymbol{x})\Big)\Big(Z(\boldsymbol{y}, \cdot) - \overline{Z}(\boldsymbol{y})\Big)\right], \ \boldsymbol{x}, \boldsymbol{y} \in D,$$

is symmetric in $\boldsymbol{x}, \boldsymbol{y}$, positive semidefinite, and continuous on $D \times D$ if continuous along 'diagonal' $\{(\boldsymbol{x}, \boldsymbol{x}) : \boldsymbol{x} \in D\}$.

The covariance operator

$$C = C_Z : L^2(D) \to L^2(D), \qquad (Cu)(\boldsymbol{x}) = \int_D u(\boldsymbol{y}) c(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}$$

is therefore selfadjoint, compact, nonnegative. Its eigenvalues $\{\lambda_m\}_{m \in \mathbb{N}}$ form a nonincreasing sequence accumulating at most at 0.

Denoting eigenfunctions by $\{Z_m\}_{m\in\mathbb{N}}$ there exists sequence of RV

$$\{\xi_m\}_{m\in\mathbb{N}} \subset L^2_{\mathbf{P}}(\Omega), \quad \mathbf{E}\left[\xi_m\right] = 0, \quad \mathbf{E}\left[\xi_k \xi_m\right] = \delta_{k,m},$$

such that the expansion

$$Z(x,\omega) = \overline{Z}(\boldsymbol{x}) + \sum_{m=1}^{\infty} \sqrt{\lambda_m}\, Z_m(\boldsymbol{x})\, \xi_m(\omega)$$

converges in $L^2_{\mathbf{P}}(\Omega; L^\infty(D))$.

[Karhunen, 1947], [Loève, 1948]

A Case Study: Radioactive Waste Disposal

Karhunen-Loève expansion

For normalized eigenfunctions $Z_m(\boldsymbol{x})$,

$$\textbf{Var}_Z(\boldsymbol{x}) := c(\boldsymbol{x},\boldsymbol{x}) = \sum_{m=1}^{\infty} \lambda_m Z_m(\boldsymbol{x})^2,$$

Total variance:

$$\int_D \textbf{Var}_Z(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \sum_{m=1}^{\infty} \lambda_m \underbrace{(Z_m, Z_m)_D}_{=1} = \operatorname{trace} C.$$

For constant variance (e.g., stationary RF),

$$\textbf{Var}_Z \equiv \sigma^2 > 0, \qquad \sum_m \lambda_m = |D|\,\sigma^2.$$

Interpretation: M first covariance eigenmodes form best rank-M approximation to C in sense of retaining maximal amount of variance.

A Case Study: Radioactive Waste Disposal

Karhunen-Loève expansion

Truncate KL expansion after M leading terms:

$$Z^{(M)}(\boldsymbol{x},\omega) = \overline{Z}(x) + \sum_{m=1}^{M} \sqrt{\lambda_m}\, Z_m(\boldsymbol{x})\, \xi_m(\omega).$$

Truncation error

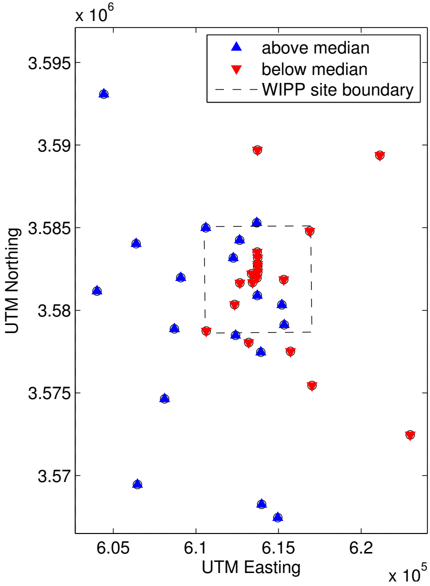
$$\mathbf{E}\left[\|Z - Z^{(M)}\|^2_{L^2(D)}\right] = \sum_{m=M+1}^{\infty} \lambda_m.$$

Choose M to retain sufficient fraction $\delta \in (0,1)$ of total variance, i.e.,

$$\frac{\mathbf{E}\left[\|Z - Z^{(M)}\|^2_{L^2(D)}\right]}{\mathbf{E}\left[\|Z\|^2_{L^2(D)}\right]} = \frac{\sum_{m=M+1}^{\infty} \lambda_m}{\sum_{m=1}^{\infty} \lambda_m} = 1 - \frac{\sum_{m=1}^{M} \lambda_m}{|D|\sigma^2} < \delta.$$

A Case Study: Radioactive Waste Disposal

WIPP Data



- transmissivity measurements at 38 test wells
- head measurements, used to obtain boundary data via statistical interpolation (kriging)
- constant layer thickness of $b = 8$m
- constant porosity of $\phi = 0.16$
- SANDIA Nat. Labs reports [Caufman et al., 1990] [La Venue et al., 1990]

A Case Study: Radioactive Waste Disposal

Probabilistic Model of Transmissivity

Merge transmissivity data with statistical model:

(1) Point estimates of parameters σ, ρ and ν via restricted maximum likelihood estimation (REML).

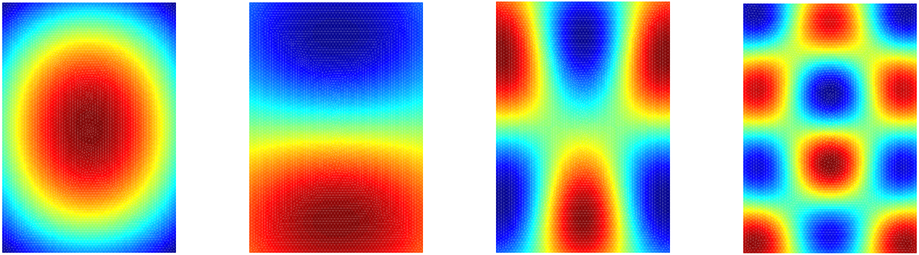
(2) Condition resulting covariance structure of $\log T$ on transmissivity measurements. (Low-rank modification of covariance operator.)

(3) Approximate $\log T$ by truncated Karhunen-Loève expansion.

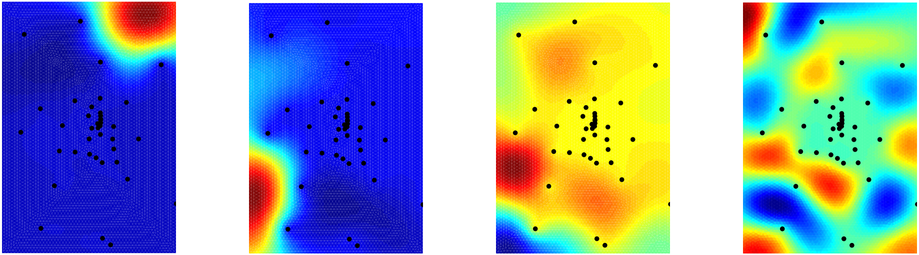
A Case Study: Radioactive Waste Disposal

WIPP KL modes conditioned on 38 transmissivity observations

unconditioned, $m = 1, 2, 9, 16$



conditioned, $m = 1, 2, 9, 16$



A Case Study: Radioactive Waste Disposal

Deterministic parametric representation

- Parametrize input RF by vector of independent Gaussian RV $\{\xi_m\}_{m=1}^M =: \boldsymbol{\xi}$.
- If ξ_m has density ρ_m and image $\Gamma_m := \xi_m(\Omega)$, then (Doob-Dynkin lemma)

$$L^2_{\mathbf{P}}(\Omega) \simeq L^2_\rho(\Gamma), \qquad \text{where} \qquad \Gamma := \times_{m=1}^{\infty} \Gamma_m, \;\; \rho = \prod_m \rho_m.$$

- Replace $Z(\boldsymbol{x},\omega)$, $p(\boldsymbol{x},\omega)$ …with $Z(\boldsymbol{x},\boldsymbol{\xi})$, $p(\boldsymbol{x},\boldsymbol{\xi})$.

BVP becomes purely deterministic with (possibly) high-dimensional parameter space:

$$\begin{aligned} \nabla\cdot[T(\boldsymbol{x},\boldsymbol{\xi})\nabla p(\boldsymbol{x},\boldsymbol{\xi})] &= 0, && \boldsymbol{x} \in D, \quad \mathbf{P}\text{-a.s.}, \\ p(\boldsymbol{x},\boldsymbol{\xi}) &= p_0(\boldsymbol{x}), && \boldsymbol{x} \in \partial D, \quad \mathbf{P}\text{-a.s.}, \end{aligned}$$

where

$$\log T(\boldsymbol{x},\boldsymbol{\xi}) = \overline{Z}(\boldsymbol{x}) + \sum_{m=1}^{M} \sqrt{\lambda_m}\, Z_m(\boldsymbol{x})\, \xi_m.$$

A Case Study: Radioactive Waste Disposal

Travel-time computations

- Generate sufficiently large ensemble of log-travel times $s(\boldsymbol{\xi}) = \log_{10} t(\boldsymbol{\xi})$
- Compute empirical CDF to quantify uncertainty in travel time.

Three sampling methods:

(1) Monte Carlo (MC) sampling of RV $\boldsymbol{\xi} \to s(\boldsymbol{\xi})$.

(N_{MC} solutions of PDE)

(2) Stochastic collocation (SC) → RF representation of velocity field $\boldsymbol{u}_{N_{SC}}(\boldsymbol{x}, \boldsymbol{\xi})$, use this to sample $s(\boldsymbol{\xi})$.

(N_{SC} solutions of PDE)

(3) Gaussian process emulator: N_{DP} MC samples of $s(\boldsymbol{\xi})$ used to calibrate surrogate of mapping $\boldsymbol{\xi} \to s(\boldsymbol{\xi})$, use this surrogate to sample $s(\boldsymbol{\xi})$.

(N_{DP} solutions of PDE)

A Case Study: Radioactive Waste Disposal

Monte Carlo Method

- Draw independent random samples $\{\boldsymbol{\xi}_j\}_{j=1}^{N_{MC}}$ of $\boldsymbol{\xi}$.
- Solve determinisic PDE for each conductivity $\exp(Z_M(\boldsymbol{x}, \boldsymbol{\xi}_j))$.
- Solve ODE for each flow field $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\xi}_j)$ and compute $s(\boldsymbol{\xi}_j)$.

A Case Study: Radioactive Waste Disposal

Monte Carlo Method

- Draw independent random samples $\{\boldsymbol{\xi}_j\}_{j=1}^{N_{MC}}$ of $\boldsymbol{\xi}$.
- Solve determinisic PDE for each conductivity $\exp(Z_M(\boldsymbol{x}, \boldsymbol{\xi}_j))$.
- Solve ODE for each flow field $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\xi}_j)$ and compute $s(\boldsymbol{\xi}_j)$.

How many samples do we need for a desired sampling error of

$$\mathbf{P}\left(\sup_{\boldsymbol{x}\in\mathbb{R}} |\hat{F}_{N_{MC}}(\boldsymbol{x}) - F(\boldsymbol{x})| \leqslant 0.01\right) \geqslant 0.95 \text{ ?}$$

Here F denotes the true CDF of s and $F_{N_{MC}}$ the empirical CDF obtained by N_{MC} samples. By Donsker's theorem we have

$$\sqrt{N_{MC}} \sup_{\boldsymbol{x}\in\mathbb{R}} |\hat{F}_{N_{MC}}(\boldsymbol{x}) - F(\boldsymbol{x})| \xrightarrow[N_{MC}\to\infty]{d} \sup_{\boldsymbol{x}\in[0,1]} |B(\boldsymbol{x})|,$$

where B is a standard Brownian Bridge on $[0, 1]$.

A Case Study: Radioactive Waste Disposal

Monte Carlo Method

- Draw independent random samples $\{\boldsymbol{\xi}_j\}_{j=1}^{N_{MC}}$ of $\boldsymbol{\xi}$.
- Solve determinisic PDE for each conductivity $\exp(Z_M(\boldsymbol{x}, \boldsymbol{\xi}_j))$.
- Solve ODE for each flow field $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\xi}_j)$ and compute $s(\boldsymbol{\xi}_j)$.

How many samples do we need for a desired sampling error of

$$\mathbf{P}\left(\sup_{\boldsymbol{x}\in\mathbb{R}} |\hat{F}_{N_{MC}}(\boldsymbol{x}) - F(\boldsymbol{x})| \leqslant 0.01\right) \geqslant 0.95 \;?$$

Here F denotes the true CDF of s and $F_{N_{MC}}$ the empirical CDF obtained by N_{MC} samples. By Donsker's theorem we have

$$\sqrt{N_{MC}} \sup_{\boldsymbol{x}\in\mathbb{R}} |\hat{F}_{N_{MC}}(\boldsymbol{x}) - F(\boldsymbol{x})| \xrightarrow[N_{MC}\to\infty]{d} \sup_{\boldsymbol{x}\in[0,1]} |B(\boldsymbol{x})|,$$

where B is a standard Brownian Bridge on $[0, 1]$.

This yields $N_{MC} \approx 20,000$. Can we do better than solving 20k PDEs?

Evaluate $v : \mathbf{\Gamma} \to V$ at collocation points $\mathbf{\Xi} := \{\boldsymbol{\xi}_j\}_{j=1}^{N_{SC}} \subset \mathbf{\Gamma}$, approximate $v_w \approx v$ in N_{SC}-dim. function space $\mathscr{V}_{\boldsymbol{\xi}}(\mathbf{\Gamma}; V)$.

Here: **Smolyak sparse tensor collocation**

$$v_w = \sum_{|\boldsymbol{i}| \leqslant w} \left[\bigotimes_{m=1}^{M} \Delta_{i_m}^{(m)} \right] v,$$

where $\Delta_0^{(m)} = 0$, $\Delta_k^{(m)} = I_k^{(m)} - I_{k-1}^{(m)}$ for $k \in \mathbb{N}$ and

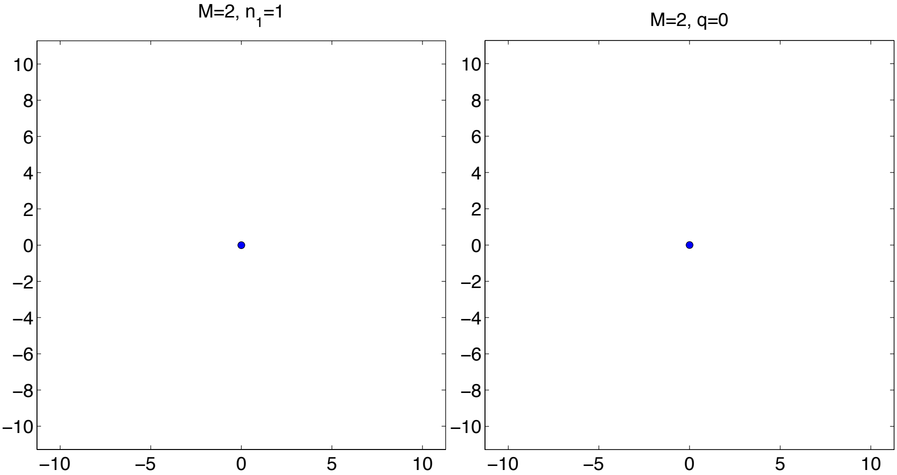
$$\left(I_k^{(m)} f \right)(\xi) := \sum_{\xi_j \in \Xi_k^{(m)}} f(\xi_j)\, \ell_j(\xi), \qquad \text{for } f : \Gamma_m \to V,$$

ℓ_j Lagrange polynomials associated with (1D) nodal sets $\Xi_k^{(m)} \subset \Gamma_m$.

Here: $\Xi_k^{(m)}$ are the $(2^{(k-1)} + 1)$th **Gauss-Hermite nodes**, $\Xi_1^{(m)} = \{0\}$.

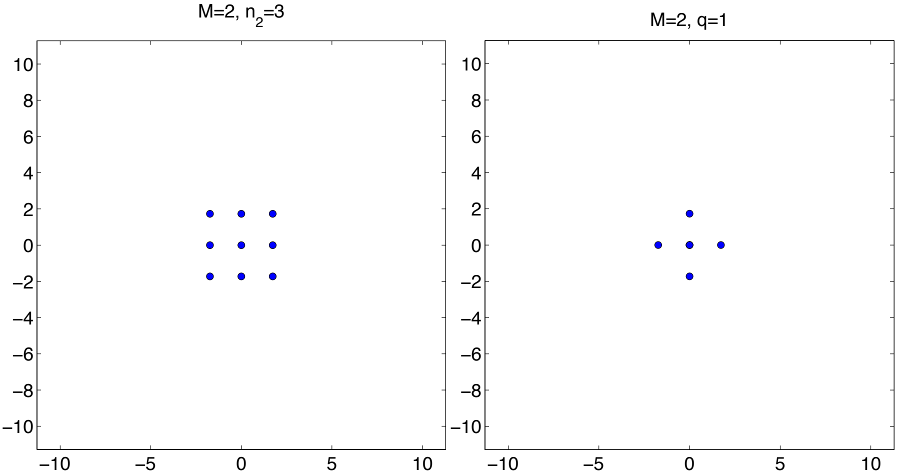
A Case Study: Radioactive Waste Disposal

Smolyak sparse grid based on Gauss-Hermite nodes



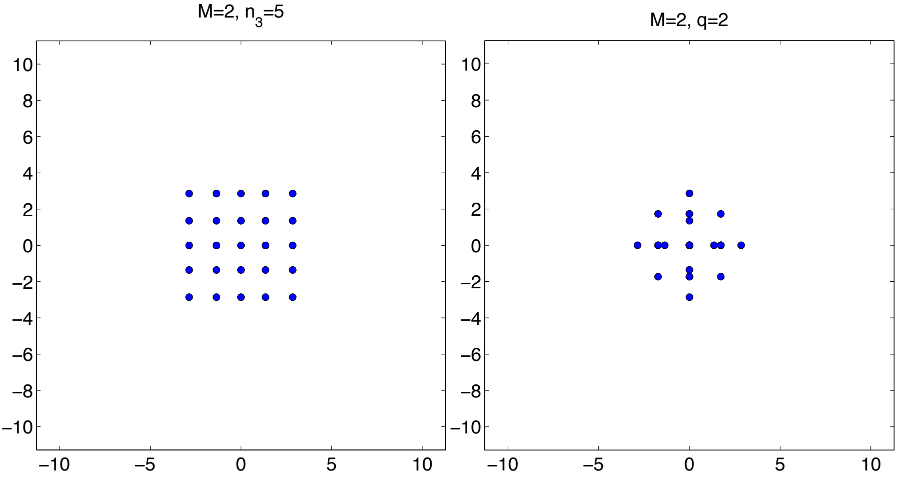
A Case Study: Radioactive Waste Disposal

Smolyak sparse grid based on Gauss-Hermite nodes



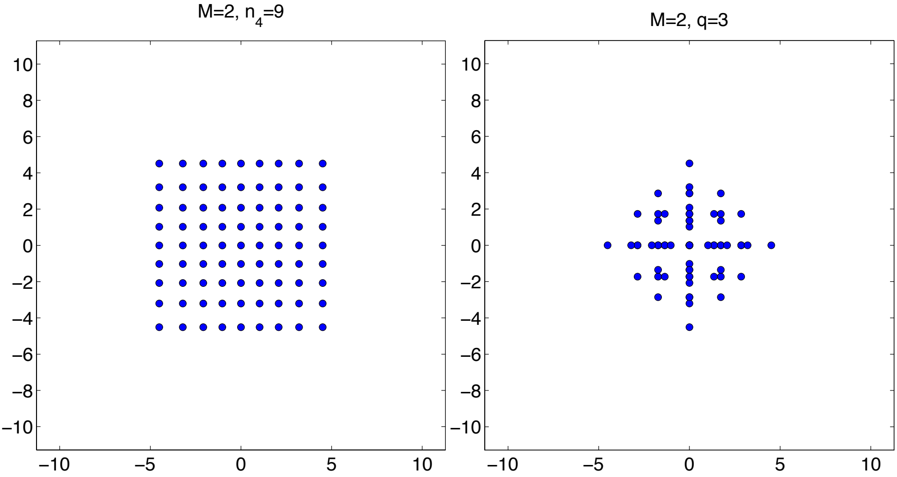
A Case Study: Radioactive Waste Disposal

Smolyak sparse grid based on Gauss-Hermite nodes



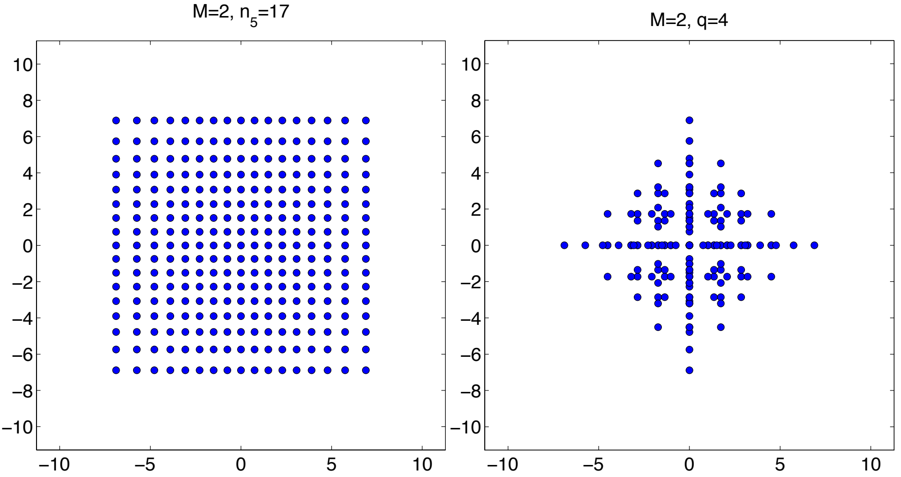
A Case Study: Radioactive Waste Disposal

Smolyak sparse grid based on Gauss-Hermite nodes



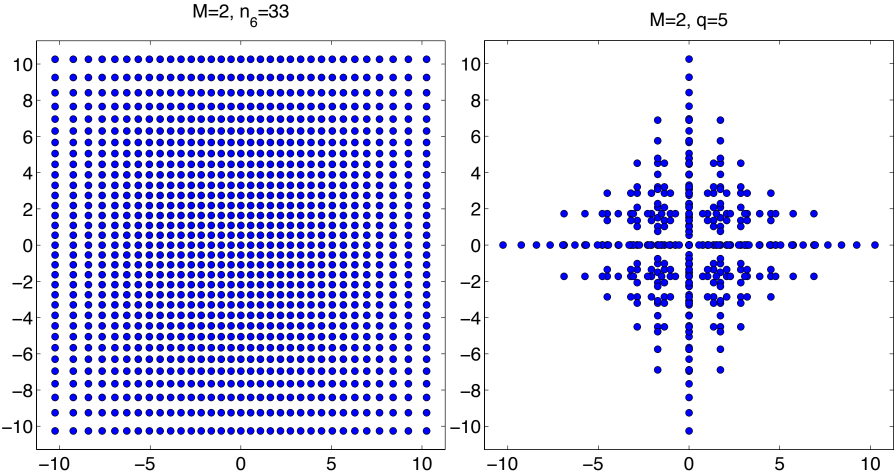
A Case Study: Radioactive Waste Disposal

Smolyak sparse grid based on Gauss-Hermite nodes



A Case Study: Radioactive Waste Disposal

Smolyak sparse grid based on Gauss-Hermite nodes



A Case Study: Radioactive Waste Disposal

An emulator is a statistical approximation to the output of a computer code

$$\boldsymbol{y} = f(\boldsymbol{x}).$$

Basic idea:

(1) Represent the code, $f(\cdot)$ as a Gaussian stochastic process.

(2) Run model for sample of design inputs $\boldsymbol{x}$ and observe outputs $\boldsymbol{y}$.

(3) Condition GP on observed outputs $\boldsymbol{y}$.

(4) Emulator provides a distribution function for the output of the computer code.

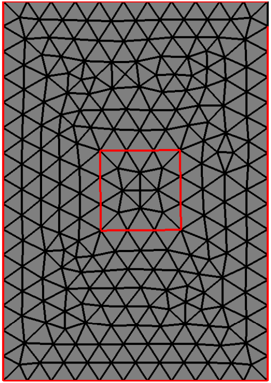
(5) Use emulator as a surrogate for computer model when performing MC analysis.

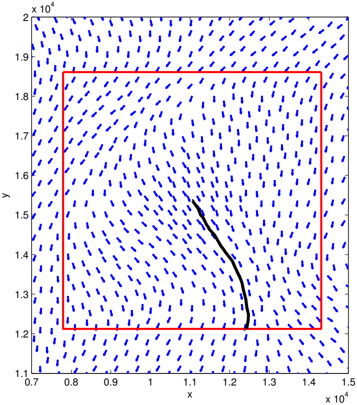
[Kennedy & O'Hagan, 2001], [Stone, 2011]

A Case Study: Radioactive Waste Disposal

Spatial Discretization

- Mixed FE discretization:
  lowest order RT elements for $\boldsymbol{u}$, pcw. constants for p.
- Fixed mesh, 29 208 triangles, (73 234 DOF)
- Flow divergence-free ⇒ discrete fluxes pcw. constant,
  (makes particle trajectory calculation trivial).



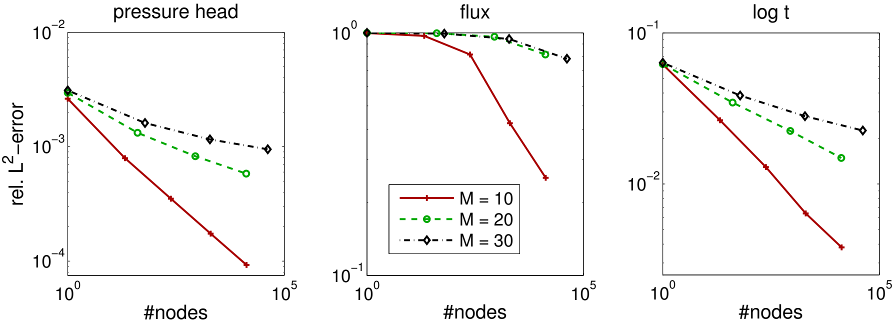


A Case Study: Radioactive Waste Disposal

Collocation error

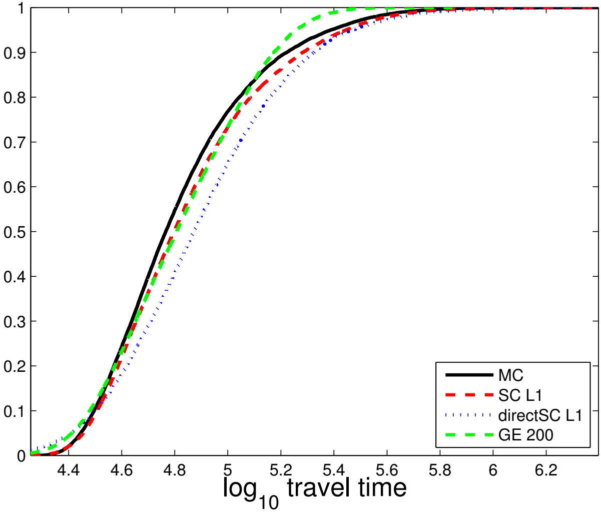
Error w.r.t. MC reference calculation with $N_{MC} = 20,000$.

Errors measured in $L^2_\rho(\mathbf{\Gamma}; L^2(D))$, $L^2_\rho(\mathbf{\Gamma}; H(\mathrm{div}, D))$ resp. $L^2_\rho(\mathbf{\Gamma}; \mathbb{R})$ norms.



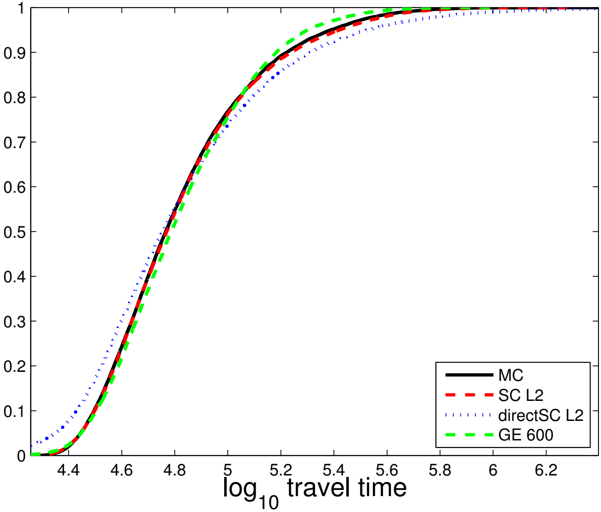
A Case Study: Radioactive Waste Disposal

Travel Time CDFs for $M = 20$ KL modes



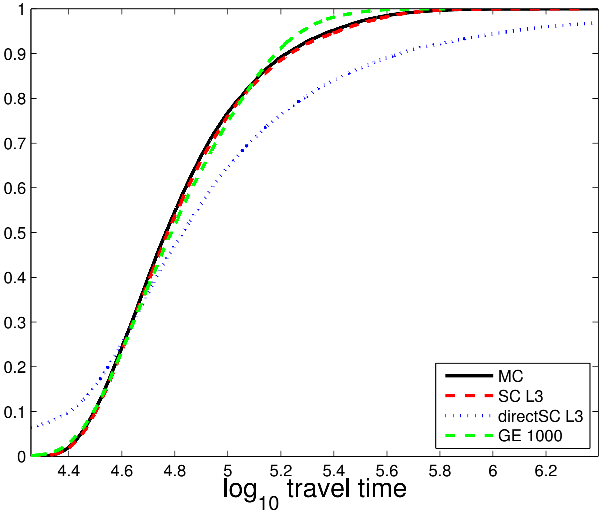
A Case Study: Radioactive Waste Disposal

Travel Time CDFs for $M = 20$ KL modes



A Case Study: Radioactive Waste Disposal

Travel Time CDFs for $M = 20$ KL modes



Kolmogorov-Smirnov Test

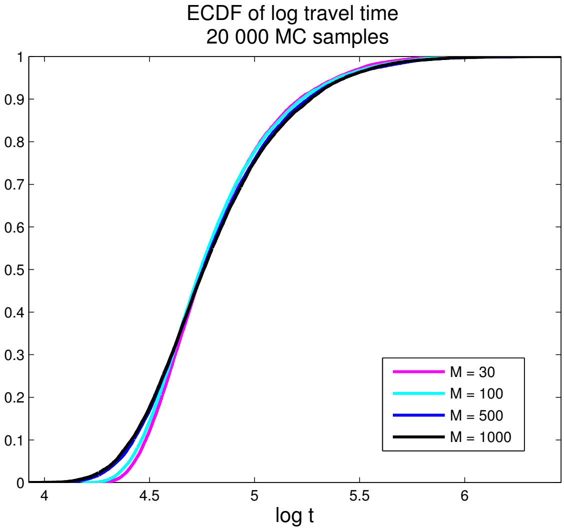
Statistical test to determine whether the random evaluations of the surrogates were drawn from the same distribution as pure MC.

Significance level: $\alpha = 0.05$. KL length: $M = 20$.

Surrogate	$N_{surrogate}$	KS-test (1K)	KS-test (20K)
SC	41	✗	✗
	881	✓	✓
	13201	✓	✓
GPE	200	✗	✗
	400	✓	✗
	600	✓	✗
	1000	✓	✗

Basically the same results for $M = 10$ and $M = 30$.

Neglected Variance



Empirical CDF of $\log t$ based on KL approximations of different length.

Contents

1 Introduction

2 Monte Carlo Methods

3 Random Fields

4 Stochastic Collocation

Contents

1. Introduction

2. Monte Carlo Methods
   - 2.1 Introduction
   - 2.2 Basic Monte Carlo Simulation
   - 2.3 Improving the Monte Carlo Method
   - 2.4 Multilevel Monte Carlo Estimators
   - 2.5 The Monte Carlo Finite Element Method

3. Random Fields

4. Stochastic Collocation

Monte Carlo Methods

Monte Carlo

Monte Carlo Methods

The Buffon Needle Problem

- George Louis Leclerc, Comte de Buffon (1707–1788), French naturalist and mathematician, posed the following problem in 1777:

  > *Let a needle of length ℓ be thrown at random onto a horizontal plane ruled with parallel straight lines spaced by a distance $d > \ell$ from each other. What is the probability p that the needle will intersect one of these lines?*

- Analog randomized experiment to approximate π, later used by Laplace.

Theorem 2.1

The probability of a needle falling in such a way that it intersects one of the lines as described above is

$$p = \frac{2\ell}{\pi d}.$$

- Let $\{H_k\}_{k\in\mathbb{N}}$ denote a sequence of i.i.d. random variables whose value is

$$H_k(\omega) = \begin{cases} 1 & \text{if } k\text{-th needle intersects a line,} \\ 0 & \text{otherwise.} \end{cases}$$

- Their common distribution is that of a Bernoulli trial with success probability $p = 2\ell/\pi d$. In particular:

$$\mathbf{E}\left[H_k\right] = p \qquad \forall k.$$

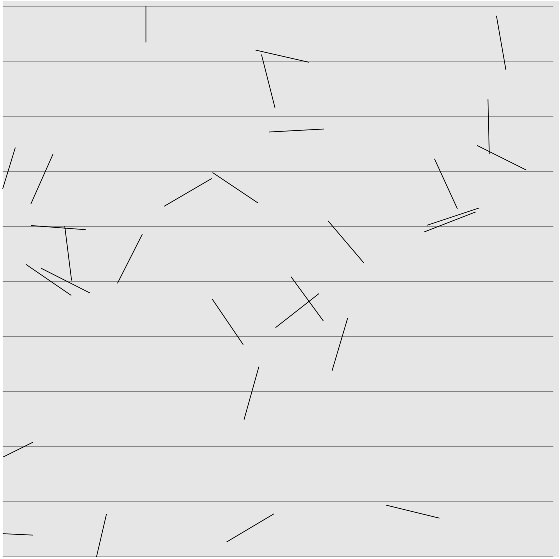
- Then $S_N = H_1 + \cdots + H_N$ is the total number of hits after N throws.
- SLLN:

$$\frac{S_N}{N} \to p \qquad \text{a.s.}$$

- Monte Carlo simulation: compute realizations of H_k by randomly sampling $X_k \sim \mathrm{U}[0, d/2]$ (distance of needle center to closest line) and $\Theta_k \sim \mathrm{U}[0, \pi/2]$ (acute angle of needle with lines) using a random number generator.

Monte Carlo Methods

The Buffon Needle Problem



- Setting $d = 2$, $\ell = 1$ gives $p = \frac{1}{\pi}$. For large N, we should have $N/S_N \approx \pi$.
- A Matlab experiment (setting `rng('default')`) yields

N	S_N	N/S_N	rel. Error
10	3	3.3	6.1e-2
100	32	3.12	5.2e-3
1000	330	3.0	3.5e-2
10000	3188	3.13	1.5e-3

- The Italian mathematician Mario Lazzarini (1901) built a machine with which to carry out many repetitions of this random experiment. His needle was 2.5 cm long and the lines 3.0 cm apart. He claims to have observed 1808 intersections for 3408 throws, corresponding to

$$\pi \approx 2 \cdot \frac{2.5}{3} \cdot \frac{3408}{1808} = 3.141592920353983\ldots$$

which corresponds to an error of $2.67 \cdot 10^{-7}$.

- Is this too good to be true?

Contents

Monte Carlo Methods

Basic Monte Carlo simulation

- Given a device for generating a sequence $\{X_k\}$ of i.i.d. realizations of a given random variable X, basic MC simulation uses the approximation

$$\mathbf{E}[X] \approx \frac{S_N}{N}, \qquad S_N = X_1 + \cdots + X_N.$$

- By the SLLN, $\frac{S_N}{N} \to \mathbf{E}[X]$ a.s.
- Similarly, for a measurable function f, $\mathbf{E}[f(X)] \approx \frac{1}{N}\sum_{k=1}^{N} f(X_k)$.
- For a RV $X \in L^2(\Omega; R)$ the standardized RV

$$X^* := \frac{X - \mathbf{E}[X]}{\sqrt{\mathbf{Var}\, X}} \quad \text{has} \quad \mathbf{E}[X^*] = 0, \ \mathbf{Var}\, X^* = 1.$$

- If $\mu = \mathbf{E}[X]$, $\sigma^2 = \mathbf{Var}\, X$, then $\mathbf{E}[S_N] = N\mu$, $\mathbf{Var}\, S_N = N\sigma^2$ and, by the CLT,

$$S_N^* = \frac{S_N - N\mu}{\sqrt{N}\sigma} \to N(0,1).$$

- Since
$$\mathbf{E}\left[\left(\frac{S_N}{N}-\mu\right)^2\right] = \mathbf{Var}\,\frac{S_N}{N} = \frac{\sigma^2}{N} \to 0,$$
we have L^2-convergence of S_N/N to μ and, by Theorem A.25, for any $\epsilon > 0$,
$$\mathbf{P}\left\{\left|\frac{S_N}{N}-\mu\right| > N^{-1/2+\epsilon}\right\} \leqslant \frac{\sigma^2}{N^{2\epsilon}}, \tag{2.1}$$
i.e., as the number N of samples increases, the probability of the error being larger than $O(N^{-1/2+\epsilon})$ converges to zero for any $\epsilon > 0$.
- If $\rho := \mathbf{E}\left[|X-\mu|^3\right] < \infty$, then the Berry-Esseen bound Theorem A.47 further gives
$$|\mathbf{P}\{S_N^* \leqslant x\} - \Phi(x)| \leqslant C\frac{\rho}{\sigma^3\sqrt{N}}, \tag{2.2}$$
where Φ denotes the cdf of $\mathrm{N}(0,1)$.

- For a RV $Z \sim \mathrm{N}(0,1)$ and $x \in \mathbb{R}$, this implies

$$\mathbf{P}(S_N^* \leqslant x) = \mathbf{P}(Z \leqslant x) + O(N^{-1/2})$$

and therefore

$$\begin{aligned}\mathbf{P}(|S_N^*| \leqslant x) &= \mathbf{P}(S_N^* \leqslant x) - \mathbf{P}(S_N^* < -x)\\ &= \mathbf{P}(Z \leqslant x) - \mathbf{P}(Z < -x) + O(N^{-1/2})\\ &= \mathbf{P}(|Z| \leqslant x) + O(N^{-1/2})\\ &= \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) + O(N^{-1/2})\end{aligned}$$

where

$$\operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) = 2\,\Phi(x) - 1.$$

- If the $O(N^{-1/2})$-term is assumed negligible, this can be used to construct (asymptotic) confidence intervals for S_N^*, i.e., the MC estimate S_N/N.

Confidence intervals from Berry-Esseen estimate

True confidence intervals are obtained if we carry along the bound in the Berry-Esseen estimate (2.2), denoted by B_N,

$$-B_N \leqslant \mathbf{P}(|S_N^*| \leqslant c) - \Phi(x) \leqslant B_N$$

i.e., for $R \geqslant 0$ we have

$$\begin{aligned}\mathbf{P}(|S_N^*| \leqslant R) &= \mathbf{P}(S_N^* \leqslant R) - \mathbf{P}(|S_N^*| < -R) \\ &\geqslant \Phi(R) - B_N - (\Phi(-R) + B_N) \\ &= \underbrace{\Phi(R) - \Phi(-R)}_{=:\gamma_R} - 2B_N\end{aligned}$$

and, in the same manner, $\mathbf{P}(|S_N^*| \leqslant R) \leqslant \gamma_R + 2B_N$, i.e.,

$$\gamma_R - 2\,B_N \leqslant \mathbf{P}\left(\mu \in \left[\frac{S_N}{N} - \frac{\sigma R}{\sqrt{N}}, \frac{S_N}{N} + \frac{\sigma R}{\sqrt{N}}\right]\right) \leqslant \gamma_R + 2\,B_N.$$

In the Buffon needle problem, we have, with RV H denoting the outcome of each needle throw,

$$\mathbf{E}\left[H\right] = p, \quad \mathbf{Var}\, H = p(1-p), \quad \mathbf{E}\left[|H-p|^3\right] = p(1-p)[1-2p+2p^2]$$

and therefore

$$S_N^* = \frac{S_N/N - p}{\sqrt{\frac{p(1-p)}{N}}} \to N(0,1).$$

Choosing $R = 2$ gives $\gamma_2 = \operatorname{erf}(\sqrt{2}) \approx 0.9545$, so that an asymptotic confidence interval of level $\gamma_2 \approx 95\%$ is obtained as

$$\left[\frac{S_N}{N} - 2\sqrt{\frac{p(1-p)}{N}}, \frac{S_N}{N} + 2\sqrt{\frac{p(1-p)}{N}}\right].$$

Monte Carlo Methods

Comparison with Lazzarini's results

In Lazzarini's experiment $\ell/d = 5/6$, $N = 3408$, giving $p = \frac{5}{3\pi} \approx 0.5305$, giving $\pi \approx \frac{5}{3} \cdot \frac{3408}{1808} = \frac{355}{113}$. This corresponds to an approximation error of

$$\left|\frac{S_N}{N} - p\right| = \left|\frac{1808}{3404} - \frac{5}{3\pi}\right| =: \epsilon_L \approx 4.5 \cdot 10^{-8}.$$

For the given values of p and N, we have

$$2\sqrt{\frac{p(1-p)}{3408}} \approx 0.0171,$$

giving a γ_2-asymptotic confidence interval around S_N/N of width

$$4\sqrt{\frac{p(1-p)}{3408}}| \approx 0.0342.$$

The γ_2 asymptotic confidence interval has a width of ϵ_L for $N > 4.9094 \cdot 10^{14}$.

To obtain true γ_R confidence intervals using the Berry-Esseen bound, note that here

$$B_N = C\frac{\rho}{\sigma^3\sqrt{N}} = C\frac{1-2p+2p^2}{\sqrt{p(1-p)N}} \leqslant \frac{0.3116}{\sqrt{N}},$$

where we have used the value $C = 0.7056$ given in [Shevtsova, 2006].

The upper bound $\gamma_R + 2B_N$ for the probability that S_N/N is within ϵ_L of the true value p after $N = 3408$ throws, corresponds to

$$\frac{\sigma}{\sqrt{N}}R \leqslant \epsilon_L, \qquad \text{i.e.,} \qquad R \leqslant R_L := \frac{\sqrt{N}\epsilon_L}{\sigma} \approx 5.2695\cdot 10^{-6}, \quad \gamma_{R_L} \approx 4.2044\cdot 10^{-6}$$

giving

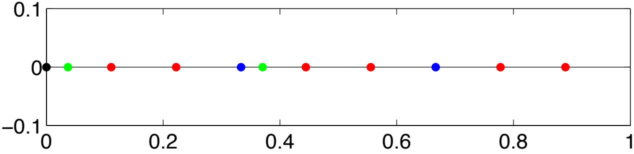
$$\mathbf{P}\left(\left|\frac{S_N}{N} - p\right| \leqslant \epsilon_L\right) \leqslant \gamma_{R_L} + 2B_{3408} \approx 0.0107.$$

Contents

In quasi-Monte Carlo methods, the samples are not chosen randomly, but special (deterministic) number sequences, known as low-discrepancy sequences, are used instead. Discrepancy is a measure of equidistribution of a number sequence.

Example: The van der Corput sequence to base 3 is such a low-discrepancy sequence for the unit interval. It is given by $x_n = \frac{k}{3^j}$, where j increases monotonically and, for each j, k runs through all nonnegative integers such that $k/3^j$ is an irreducible fraction. The ordering in k is obtained by representing k in base 3 and reversing the digits. The first 11 numbers are

$$\{x_n\}_{n=1}^{11} = 0, \frac{1}{3}, \frac{2}{3}, \frac{1}{9}, \frac{4}{9}, \frac{7}{9}, \frac{2}{9}, \frac{5}{9}, \frac{8}{9}, \frac{1}{27}, \frac{10}{27}.$$



- Replacing i.i.d. random numbers sampled from $\mathrm{U}[0,1]$ in a standard Monte Carlo approximation of $\mathbf{E}\left[f(X)\right]$ for some $f \in C^\infty(0,1)$ and $X \sim \mathrm{U}[0,1]$, by the van der Corput sequence of length N, yields a quasi-Monte Carlo method.
- The convergence rate is improved from $O(N^{-1/2})$ to $O(N^{-2})$.
- Although this improvement is impressive, the method does not generalise easily and the rate of convergence depends on the problem.
- In particular, the rate of convergence for a quasi-Monte Carlo method generally does depend on the dimension.

Monte Carlo Methods

Variance reduction

The constant in the MC convergence rate appearing in (2.1) is the variance of the RV from which MC samples are being drawn. By designing an equivalent MC approximation with lower variance, we can expect to obtain faster convergence.

- To approximate $\mathbf{E}[X]$ by standard MC, we draw independent samples $\{X_k\}_{k=1}^N$ of X and compute the sample average S_N/N.
- Now assume a second set of samples $\tilde{X}_k$ of X is given with sample average $\tilde{S}_N/N$.
- Since both sample averages converge to $\mathbf{E}[X]$, so does $\frac{1}{2}(S_N/N + \tilde{S}_N/N)$.
- When X_k and $\tilde{X}_k$ are negatively correlated they are called antithetic samples, and the approximation $\frac{1}{2N}(S_N + \tilde{S}_N)$ is a more reliable approximation of $\mathbf{E}[X]$ than $\frac{1}{2N}S_{2N}$.

Theorem 2.2

Let thew two sequences $\{X_k\}$ and $\{\tilde{X}_k\}$ of random variables be identically distributed with

$$\textbf{Cov}(X_j, X_k) = \textbf{Cov}(\tilde{X}_j, \tilde{X}_k) = 0 \quad \text{for } j \neq k.$$

Then the sample averages S_N/N and $\tilde{S}_N/N$ satisfy

$$\textbf{Var}\,\frac{S_N + \tilde{S}_N}{2N} = \textbf{Var}\,\frac{S_{2N}}{2N} + \frac{1}{2}\,\textbf{Cov}\left(\frac{S_N}{N}, \frac{\tilde{S}_N}{N}\right) \leqslant \textbf{Var}\,\frac{S_N}{N}. \tag{2.3}$$

- Worst case: Variance of average of N samples and N antithetic samples less than variance of N independent samples.
- Best case: negatively correlated S_N/N and $\tilde{S}_N/N$, therefore variance of N samples and N antithetic samples less than variance of $2N$ indepependent samples.

Monte Carlo Methods

Example: Predator-prey dynamical system

Consider the popular model of the dynamics of two interacting populations

$$\dot{\boldsymbol{u}} = \begin{bmatrix} \dot{u}_1 \\ \dot{u}_2 \end{bmatrix} = \begin{bmatrix} u_1(1-u_2) \\ u_2(u_1-1) \end{bmatrix}, \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0.$$

Assume the vector of initial conditions $\boldsymbol{u}_0$ is uncertain and that it is modeled as a random vector $\boldsymbol{u}_0 \sim \mathrm{U}(\Gamma)$, where Γ denotes the square

$$\Gamma = \overline{\boldsymbol{u}}_0 + [-\epsilon, \epsilon]^2.$$

- **Goal:** estimate $\mathbf{E}[u_1(T)]$ at time $T > 0$.
- Denote by $\boldsymbol{u}_n = \boldsymbol{u}_n(\omega)$ the explicit Euler approximation after n time steps of length Δt starting with initial data $\boldsymbol{u}_0 = \boldsymbol{u}_0(\omega)$.
- Define $\phi(\boldsymbol{u}) = u_1$ for $\boldsymbol{u} = [u_1, u_2]^{\mathsf{T}} \in \mathbb{R}^2$, estimate $\mathbf{E}[\phi(\boldsymbol{u}_{n_T})]$ for $n_T \Delta t = T$, using the MC method.
- Denote by $\overline{S}_N := S_N/N$ the average over N samples of $u_1(T)$.
- Expect better approximations for N large and Δt small.

Monte Carlo Methods

Example: Predator-prey dynamical system

- **Notation:**

forward map:	$G : \Gamma \to C([0,T];\mathbb{R}^2)$
discretized forward map:	$G_{\Delta t} : \Gamma \to C([0,T];\mathbb{R}^2)$
quantity of interest (QoI):	$Q : C([0,T];\mathbb{R}^2) \to \mathbb{R}, \quad \boldsymbol{u} \mapsto u_1(T) = \phi(\boldsymbol{u}(T))$
approximation of QoI:	$Q_{\Delta t} := \phi(\boldsymbol{u}_{n_T}) = \phi(G_{\Delta t}(\boldsymbol{u}_0)\vert_{t=T})$
MC estimate, N samples:	$\widehat{Q}_{\Delta t} := \widehat{Q}_{\Delta t,N} \approx \mathbf{E}\left[Q_{\Delta t}\right] \approx \mathbf{E}\left[Q\right].$

- Error with N samples and $n_T = T/\Delta t$ time steps:

$$e_{N,\Delta t} = |\mathbf{E}\left[Q\right] - \widehat{Q}_{\Delta t}| \leqslant \underbrace{|\mathbf{E}\left[Q\right] - \mathbf{E}\left[Q_{\Delta t}\right]|}_{\text{explicit Euler error}} + \underbrace{|\mathbf{E}\left[Q_{\Delta_t}\right] - \widehat{Q}_{\Delta t}|}_{\text{Monte Carlo error}}$$

- Explicit Euler error:
$$\|\boldsymbol{u}(T) - \boldsymbol{u}^{\Delta t}(T)\| \leqslant K\Delta t.$$
- ϕ Lipschitz-continuous with constant $L = 1$:
$$|\phi(\boldsymbol{u}(T)) - \phi(\boldsymbol{u}^{\Delta t}(T))| \leqslant K\,L\,\Delta t.$$
- Therefore
$$|\mathbf{E}\left[Q\right] - \mathbf{E}\left[Q_{\Delta t}\right]| = |\mathbf{E}\left[Q - Q_{\Delta t}\right]| \leqslant K\,L\,\Delta t. \tag{2.4}$$

- For MC error, apply CLT, confidence intervals: if $\mathbf{Var}\, Q_{\Delta t} = \sigma^2$,

$$\mathbf{P}\left(\left|\mathbf{E}\left[Q_{\Delta t}\right] - \widehat{Q}_{\Delta t,N}\right| \leqslant \frac{2\sigma}{\sqrt{N}}\right) > \gamma_2 + O(N^{-1/2})$$

- Combined with (2.4):

$$\mathbf{P}\left(e_{N,\Delta t} \leqslant K\, L\, \Delta t + \frac{2\sigma}{\sqrt{N}}\right) > \gamma_2 + O(N^{-1/2}).$$

- Balance discretization and MC errors:

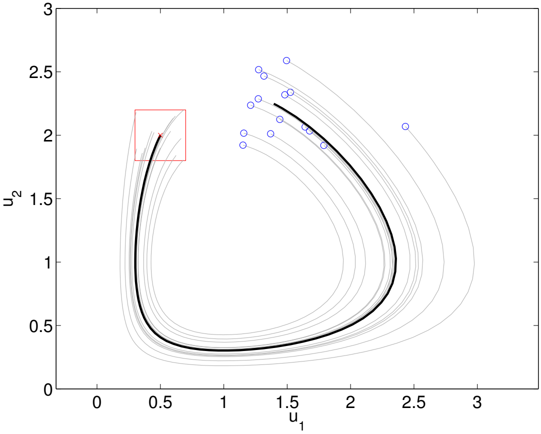
$$KL\Delta t \approx \frac{\delta}{2}, \quad \frac{2\sigma}{\sqrt{N}} \approx \frac{\delta}{2},$$

leads to

$$\Delta t \approx \frac{\delta}{2KL} \quad \text{and} \quad N \approx \frac{16\sigma^2}{\delta^2}.$$

Monte Carlo Methods

Example: Predator-prey dynamical system



Population dynamics problem integrated over $[0, T = 6]$ with $\boldsymbol{u}_0 = [0.5, 2] + \mathrm{U}[-\epsilon, \epsilon]$ for $\epsilon = 0.2$. Unperturbed trajectory (black) along with 15 perturbed trajectories. For the unperturbed trajectory $u_1(T) = 1.3942$.

Monte Carlo Methods

Example: Predator-prey dynamical system, antithetic sampling

We may introduce antithetic sampling to this problem by noting that, if $\boldsymbol{u}_0 \sim \mathrm{U}(\Gamma)$, then the same holds for the random vector

$$\tilde{\boldsymbol{u}}_0 := 2\overline{\boldsymbol{u}}_0 - \boldsymbol{u}_0.$$

Thus, the trajectories generated by the random initial data $\tilde{\boldsymbol{u}}_0$ have the same distribution as those generated by $\boldsymbol{u}_0$.

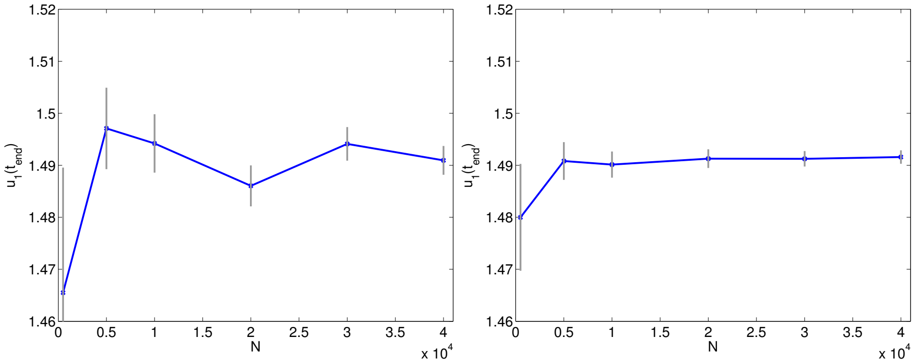
- Denote by $X_k = \phi(\boldsymbol{u}^{\Delta t}(T))$ the basic samples, by $\tilde{X}_k$ the antithetic counterparts. Note that all pairs of samples are independent except each sample and its antithetic counterpart.
- We estimate $\overline{S}_{2N}$ using the sample variance.
- To estimate $\frac{1}{2}\left(\overline{S}_N + \overline{\tilde{S}}_N\right)$ by (2.3), note that

$$\mathbf{Cov}(\overline{S}_N, \overline{\tilde{S}}_N) = \frac{1}{N^2}\,\mathbf{Cov}(S_N, \tilde{S}_N) = \frac{1}{N^2}\sum_{k=1}^{N}\mathbf{Cov}(X_k, \tilde{X}_k) = \frac{1}{N}\,\mathbf{Cov}(X, \tilde{X})$$

  The last quantity can be estimated using the sample covariance.

Monte Carlo Methods

Example: Predator-prey dynamical system



MC estimation of $\mathbf{E}\left[u_1(T)\right]$ using standard MC with N samples (left) versus MC with antithetic sampling using $N/2$ samples (right) of the initial data. Both curves show the estimate along with a 95% (asymptotic) confidence interval.

Contents

Multilevel Monte Carlo Methods

Discretization

The following summary of basic MLMC techniques and analysis closely follows [Teckentrup, 2013], see also [Cliffe et al., 2011]

To estimate the expectation $\mathbf{E}[Q]$ of a (random) quantity of interest (QoI) Q, assume only approximations $Q_h \approx Q$ are computable, where $h > 0$ denotes a discretization parameter for which

$$\lim_{h\to 0} \mathbf{E}[Q_h] = \mathbf{E}[Q].$$

More precisely, we shall assume the error in mean to converge at a rate α, i.e.,

$$|\mathbf{E}[Q_h - Q]| \lesssim h^\alpha, \qquad \text{as } h \to 0, \qquad \alpha > 0.$$

Multilevel Monte Carlo Methods

Mean-square error

- Given an unbiased estimator $\widehat{Q}_h$ for $\mathbf{E}[Q_h]$, the associated mean-square error (MSE) may always be decomposed as

$$\begin{aligned}\mathbf{E}\left[\left(\widehat{Q}_h - \mathbf{E}[Q]\right)^2\right] &= \mathbf{E}\left[\left(\widehat{Q}_h - \mathbf{E}\left[\widehat{Q}_h\right] + \mathbf{E}\left[\widehat{Q}_h\right] - \mathbf{E}[Q]\right)^2\right] \\ &= \mathbf{E}\left[\left(\widehat{Q}_h - \mathbf{E}\left[\widehat{Q}_h\right]\right)^2\right] + \left(\mathbf{E}\left[\widehat{Q}_h\right] - \mathbf{E}[Q]\right)^2 \\ &= \mathbf{Var}\,\widehat{Q}_h + \left(\mathbf{E}[Q_h] - \mathbf{E}[Q]\right)^2 \\ &= \mathbf{Var}\,\widehat{Q}_h + \mathbf{E}[Q_h - Q]^2\end{aligned}$$

  consisting of the variance of the estimator and the squared expectation of the discretization error (systematic error, bias).
- We shall sometimes refer to the root mean-square error (RMSE), which is simply the square root of the MSE, i.e., the L^2-norm of the estimation error

$$\sqrt{\mathbf{E}\left[\left(\widehat{Q}_h - \mathbf{E}[Q]\right)^2\right]}$$

Multilevel Monte Carlo Methods

Standard MC estimator

- If the standard Monte Carlo estimator $\widehat{Q}_h = \widehat{Q}_{h,N}^{\mathsf{MC}}$ with N samples is used, and $Q_h^{(i)}$ denote i.i.d. RV with the same distribution as Q_h, then

  $$\mathbf{Var}\,\widehat{Q}_{h,N}^{\mathsf{MC}} = \mathbf{Var}\left(\frac{1}{N}\sum_{i=1}^{N} Q_h^{(i)}\right) = \frac{1}{N^2} N\,\mathbf{Var}\,Q_h = \frac{\mathbf{Var}\,Q_h}{N},$$

  giving

  $$\mathbf{E}\left[\left(\widehat{Q}_{h,N}^{\mathsf{MC}} - \mathbf{E}\left[Q\right]\right)^2\right] = \frac{\mathbf{Var}\,Q_h}{N} + \mathbf{E}\left[Q_h - Q\right]^2.$$

- We denote by $\mathscr{C}(\widehat{Q})$ the cost, in terms of the number of floating-point operations required for its evaluation, associated with an estimator $\widehat{Q}$. The cost will often depend on the type of discretization, typically inversely proportional to h or, more generally, satisfying a relation of the form

  $$\mathscr{C}(Q_h^{(i)}) \lesssim h^{-\gamma}, \qquad \gamma > 0.$$

  so that $\mathscr{C}(\widehat{Q}_{h,N}^{\mathsf{MC}}) \lesssim N h^{-\gamma}$.

Multilevel Monte Carlo Methods

Cost scaling

- To balance the two error components, assume each ist bounded by $\frac{\epsilon^2}{2}$, resulting in a total bound of ϵ for the RMSE.
- Assuming $\mathbf{Var}\, Q_h$ is approximately constant independent of h, this error balance requires

$$N \gtrsim \epsilon^{-2} \qquad \text{and} \qquad h \lesssim \epsilon^{1/\alpha}.$$

- Since the cost per sample was assumed to satisfy $\mathscr{C}(Q_h^{(i)}) \lesssim h^{-\gamma}$, this gives

$$\mathscr{C}(\widehat{Q}_{h,N}^{\mathsf{MC}}) \lesssim N h^{-\gamma},$$

  whereby the total cost of achieving a RMSE of $O(\epsilon)$ using a standard MC estimator is

$$\mathscr{C}_\epsilon(\widehat{Q}_{h,N}^{\mathsf{MC}}) \lesssim \epsilon^{-2-\gamma/\alpha}.$$

Multilevel Monte Carlo Methods

Multilevel estimator

- The idea underlying multilevel estimators is to use realizations of Q_h on different levels, i.e., for different values $h_0, \dots, h_L$ of the discretization parameter, and decompose $\mathbf{E}[Q_h]$ as

$$\mathbf{E}[Q_h] = \mathbf{E}[Q_{h_0}] + \sum_{\ell=1}^{L} \mathbf{E}\left[Q_{h_\ell} - Q_{h_{\ell-1}}\right] =: \sum_{\ell=0}^{L} \mathbf{E}[Y_\ell],$$

where

$$h_\ell = s^{-1} h_{\ell-1}, \qquad h_0 > 0, \quad \ell = 1, \dots, L, \quad s \in \mathbb{N}\backslash\{1\}. \tag{2.5}$$

- Given (unbiased) estimators $\{\widehat{Y}_\ell\}_{\ell=0}^{L}$ for $\mathbf{E}[Y_\ell]$, we refer to

$$\widehat{Q}_h^{\mathsf{ML}} := \sum_{\ell=0}^{L} \widehat{Y}_\ell$$

as a multilevel estimator for Q.

- Since all expectations $\mathbf{E}[Y_\ell]$ are sampled independently, we have

$$\mathbf{Var}\, \widehat{Q}_h^{\mathsf{ML}} = \sum_{\ell=0}^{L} \mathbf{Var}\, \widehat{Y}_\ell.$$

- If each $\widehat{Y}_\ell$ is itself a standard Monte Carlo estimator, i.e.,

  $$\widehat{Y}_0 = \widehat{Y}^{\mathsf{MC}}_{0,N_0} := \widehat{Q}^{\mathsf{MC}}_{h_0,N_0}$$

  and

  $$\widehat{Y}^{\mathsf{MC}}_{\ell,N_\ell} := \frac{1}{N_\ell}\sum_{i=0}^{N_\ell}\left(Q^{(i)}_{h_\ell} - Q^{(i)}_{h_{\ell-1}}\right), \qquad \ell = 1,\dots,L,$$

  one obtains a multilevel Monte Carlo estimator, denoted $\widehat{Q}^{\mathsf{MLMC}}_{h,\{N_\ell\}}$.
- The associated MSE then has the standard decomposition

  $$\mathbf{E}\left[\left(\widehat{Q}^{\mathsf{MLMC}}_{h,\{N_\ell\}} - \mathbf{E}\left[Q\right]\right)^2\right] = \sum_{\ell=0}^{L}\frac{\mathbf{Var}\,Y_\ell}{N_\ell} + \mathbf{E}\left[Q_h - Q\right]^2 \tag{2.6}$$

  into estimation variance and bias.

Multilevel Monte Carlo Methods

MLMC scaling

- To achieve a balanced RMSE of ϵ, note that the the bias term in (2.6) is the same as for the standard MC estimator, leading again to a choise of $h = h_L$ satisfying $h \lesssim \epsilon^{1/\alpha}$.
- Achieving a bound of $\epsilon^2/2$ for the variance term in the MSE is typically possible at lower cost than for standard MC for the following two reasons:
- If $Q_h \to Q$ also in mean square, then $\mathbf{Var}\, Y_\ell = \mathbf{Var}(Q_{h_\ell} - Q_{h_{\ell-1}}) \to 0$ as $\ell \to \infty$, allowing for smaller and smaller sample sizes N_ℓ on finer and finer levels.
- As $\epsilon \to 0$, the discretization parameter h_0 on the coarsest level can remain fixed, leading to fixed cost per sample there.

Multilevel Monte Carlo Methods

MLMC cost

- The cost of the MLMC estimator is

$$\mathscr{C}(\widehat{Q}^{\mathsf{MLMC}}_{h,\{N_\ell\}}) = \sum_{\ell=0}^{L} N_\ell \mathscr{C}_\ell, \qquad \mathscr{C}_\ell := \mathscr{C}(Y_\ell^{(i)}).$$

- Treating the N_ℓ as continuous variables, the variance of the MLMC estimator is minimized for a fixed cost for

$$N_\ell \simeq \sqrt{\frac{\mathbf{Var}\, Y_\ell}{\mathscr{C}_\ell}} \tag{2.7}$$

  with the implied constant chosen to make the total variance equal to $\epsilon^2/2$.

- This results in a total cost on level ℓ proportional to $\sqrt{\mathscr{C}_\ell \, \mathbf{Var}\, Y_\ell}$ and therefore

$$\mathscr{C}(\widehat{Q}^{\mathsf{MLMC}}_{h,\{N_\ell\}}) \lesssim \sum_{\ell=0}^{L} \sqrt{\mathscr{C}_\ell \, \mathbf{Var}\, Y_\ell}$$

Multilevel Monte Carlo Methods

MLMC cost

- If $\mathbf{Var}\, Y_\ell$ decays faster than $\mathscr{C}_\ell$ increases, the cost on level $\ell = 0$ dominates, and, since $N_0 \eqsim \epsilon^{-2}$, the cost ratio of MLMC to ML estimation is approximately
$$\frac{\mathscr{C}_0}{\mathscr{C}_L} \eqsim \left(\frac{h_L}{h_0}\right)^{\gamma}.$$
- If $\mathscr{C}_\ell$ increases faster than $\mathbf{Var}\, Y_\ell$ decays, then the cost on level $\ell = L$ dominates, and then the cost ratio is approximately
$$\frac{\mathbf{Var}\, Y_L}{\mathbf{Var}\, Y_0},$$
which is $O(\epsilon^2)$ if h_0 is such that $\mathbf{Var}\, Y_0 \eqsim \mathbf{Var}\, Q_{h_0}$.

Multilevel Monte Carlo Methods

Theorem 2.3

Let $\{h_\ell\}_{\ell=0}^L$ satisfy (2.5), $\epsilon < \exp(-1)$, and assume there exist constants $\alpha, \beta, \gamma, \delta, c_{\mathsf{M1}}, c_{\mathsf{M2}}, c_{\mathsf{M4}} > 0$ such that $\alpha \geqslant \min\{\beta, \gamma/\delta\}$ and $\delta \in (\frac{1}{2}, 1]$.
Assume further that

(M1) $|\mathbf{E}[Q_{h_\ell}] - \mathbf{E}[Q]| \leqslant c_{\mathsf{M1}} h_\ell^\alpha$.

(M2) $\mathbf{Var}\, \widehat{Y}_\ell \leqslant c_{\mathsf{M2}} N_\ell^{-1/\delta} h_\ell^\beta$.

(M3) $\mathbf{E}\left[\widehat{Y}_\ell\right] = \begin{cases} \mathbf{E}[Q_{h_0}], & \ell = 0, \\ \mathbf{E}\left[Q_{h_\ell} - Q_{h_{\ell-1}}\right], & \ell = 1, \dots, L. \end{cases}$

(M4) $\mathscr{C}(\widehat{Y}_\ell) \leqslant N_\ell h_\ell^{-\gamma}$.

Then there exists $\{N_\ell\}_{\ell=0}^L$ such that $\mathbf{E}\left[\left(\widehat{Q}_h^{\mathsf{ML}} - \mathbf{E}[Q]\right)^2\right] \leqslant \epsilon^2$ where $h = h_L$ and

$$\mathscr{C}(\widehat{Q}_h^{\mathsf{ML}}) \leqslant c \begin{cases} \epsilon^{-2\delta}, & \text{if } \delta\beta > \gamma, \\ \epsilon^{-2\delta} |\log \epsilon|^{1+\delta}, & \text{if } \delta\beta = \gamma, \\ \epsilon^{-2\delta-(\gamma-\delta\beta)/\alpha}, & \text{if } \delta\beta < \gamma, \end{cases}$$

where the constant c depends on $c_{\mathsf{M1}}, c_{\mathsf{M2}}$ and c_{M4}.

- The following MLMC algorithm computes the optimal values of N_ℓ ‘on the fly’ using (unbiased) sample averages and sample variances of Y_ℓ.
- We assume there exists an $h^\star > 0$ such that the error decay in $|\mathbf{E}[Q_h - Q]|$ is monotonic for $h \leqslant h^\star$ and satisfies $|\mathbf{E}[Q_h - Q]| \eqsim h^\alpha$.
- This ensures that $|\mathbf{E}[Y_L]| \eqsim h^\alpha$ since $s > 1$ and thus $|\widehat{Y}_L| \eqsim h^\alpha$ for N_L sufficiently large.
- This gives a computable error estimator to determine whether h is sufficiently small or whether L needs to be increased.

Multilevel Monte Carlo Methods

MLMC Algorithm

Algorithm 1: MLMC algorithm

```
1 $L \leftarrow 0$.
2 Estimate $\mathbf{Var}\, Y_L$ by the sample variance of an initial number of samples.
3 Calculate optimal $\{N_\ell\}_{\ell=1}^{L}$ using (2.7).
4 Evaluate extra samples at each level as needed for the new $N_\ell$.
5 if $L \geqslant 1$ then
6     test for convergence using $\widehat{Y}_L \approx h^\alpha$.
7 if not converged or $L = 0$ then
8     $L \leftarrow L + 1$ and go back to 2.
```

- Step 3 aims to make the variance of the MLMC estimator less than $\epsilon^2/2$.
- Step 5 ensures that the remaining bias is less than $\epsilon/\sqrt{2}$.

Contents

1 Introduction

2 Monte Carlo Methods

2.1 Introduction

2.2 Basic Monte Carlo Simulation

2.3 Improving the Monte Carlo Method

2.4 Multilevel Monte Carlo Estimators

2.5 The Monte Carlo Finite Element Method

3 Random Fields

4 Stochastic Collocation

Monte Carlo Finite Element Method

We return to our model elliptic boundary value problem with random data

$$-\nabla\cdot(a\nabla u) = f, \qquad \text{on } D \subset \mathbb{R}^2, \qquad u_{|\partial D} = 0, \tag{2.8}$$

where a and f are random fields defined on D with respect to a probability space $(\Omega, \mathfrak{A}, \mathbf{P})$.

- If f is random, we assume $f(\cdot,\omega) \in L^2(D)$ for (almost) all $\omega \in \Omega$.
- Our goal is to use the MC method to estimate a quantity of interest which depends on the (random) solution u. We focus, for now, on the mean $\mathbf{E}\left[u(\boldsymbol{x},\cdot)\right]$ and variance $\mathbf{Var}\, u(\boldsymbol{x},\cdot)$.
- With each of N i.i.d. realizations $a^{(j)} = a(\cdot,\omega_j)$ and $f^{(j)} = f(\cdot,\omega_j)$ we associate the unique solution $u^{(j)}$, approximate $u_h^{(j)} \approx u^{(j)}$ using the finite element method and compute the ($H_0^1(D)$-valued) estimates

$$\mu_{N,h} := \frac{1}{N}\sum_{j=1}^{N} u_h^{(j)}, \qquad \sigma_{N,h}^2 := \frac{1}{N-1}\sum_{j=1}^{N}\left(u_h^{(j)} - \mu_{N,h}\right)^2$$

Monte Carlo Finite Element Method

Assumptions on a

To ensure a unique solution $u^{(j)}$ for each realization, we could require the coefficient a to satisfy Assumption B.3. However, this proves too restrictive in many applications, and for many cases it is sufficient to require merely realization-wise bounds:

Assumption 2.4

For almost all $\omega \in \Omega$, realizations $a(\cdot, \omega)$ of the coefficient function $a = a(\boldsymbol{x})$ lie in $L^\infty(D)$ and satisfy

$$0 < a_{\min}(\omega) \leqslant a(\boldsymbol{x}, \omega) \leqslant a_{\max}(\omega) < \infty \qquad \text{a.e. in } D, \tag{2.9}$$

where

$$a_{\min}(\omega) := \operatorname*{ess\,inf}_{\boldsymbol{x} \in D} a(\boldsymbol{x}, \omega), \qquad a_{\max}(\omega) := \operatorname*{ess\,sup}_{\boldsymbol{x} \in D} a(\boldsymbol{x}, \omega). \tag{2.10}$$

Monte Carlo Finite Element Method

Realization-wise solution

For any realization ω for which Assumption 2.4 holds and $f(\omega) \in L^2(D)$, we may apply the Lax-Milgram lemma and obtain a unique solution of (2.8).

Theorem 2.5

Let Assumption 2.4 hold and $f(\cdot,\omega) \in L^2(D)$ **P**-a.s. Then (2.8) has a unique solution $u(\cdot,\omega) \in H_0^1(D)$ **P**-a.s.

The following theorem provides sufficient conditions for the realization-wise solutions u to have finite p-th moments, i.e., to lie in $L^p(\Omega; H_0^1(D))$.

Monte Carlo Finite Element Method

Realization-wise summability

Theorem 2.6

Under Assumption 2.4, assume the mappings $a : \Omega \to L^\infty(D)$ and $f : \Omega \to L^2(D)$ are measurable, let $V^h \subset H_0^1(D)$ denote a closed subspace and $u_h : \Omega \to V^h$ satisfy **P**-a.s.

$$\int_D a(\boldsymbol{x},\omega)\nabla u_h(\boldsymbol{x},\omega) \cdot \nabla v(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \int_D f(\boldsymbol{x},\omega) v(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \qquad \forall v \in V^h.$$

Then, with C_D the Poincaré-Friedrichs constant from Lemma B.7:

(a) If $f \in L^2(D)$ is deterministic, then $1/a_{\min} \in L^p(\Omega;\mathbb{R})$ with $p \geqslant 1$ implies

$$\|u_h\|_{L^p(\Omega;H_0^1(D))} \leqslant C_D \|a_{\min}^{-1}\|_{L^p(\Omega;\mathbb{R})} \|f\|_{L^2(D)}.$$

(b) If $1/a_{\min} \in L^q(\Omega;\mathbb{R})$ and $f \in L^r(\Omega;L^2(D))$ with $q, r \geqslant 1$, $1/q + 1/r = 1/p \leqslant 1$, then

$$\|u_h\|_{L^p(\Omega;H_0^1(D))} \leqslant C_D \|a_{\min}^{-1}\|_{L^q(\Omega;\mathbb{R})} \|f\|_{L^r(\Omega;L^2(D))}.$$

If, in addition, a and f are independent, the above bound holds with $q = r = p$.

Monte Carlo Finite Element Method

Mean finite element error

Assumption 2.7

There exists a constant $K_2 > 0$ such that, for every $f \in L^2(D)$, we have $u \in L^4(\Omega; H^2(D))$ and

$$|u|_{L^4(\Omega;H^2(D))} \leqslant K_2 \|f\|_{L^2(\Omega;L^2(D))}.$$

Theorem 2.8

Under the conditions of Theorem 2.5 together with Assumption 2.7 and assuming that $a_{\min}^{-1/2} a_{\max}^{1/2} \in L^4(\Omega; \mathbb{R})$, the piecewise linear finite element approximation u_h with respect to a shape-regular triangulation $\mathscr{T}_h$ satisfies

$$\|u - u_h\|_{L^2(\Omega;H_0^1(D))} \leqslant Kh \|a_{\min}^{-1/2} a_{\max}^{1/2}\|_{L^4(\Omega;\mathbb{R})} \|f\|_{L^2(\Omega)}.$$

Monte Carlo Finite Element Method

Error analysis

We split the error in approximating $\mathbf{E}[u]$ by the MC estimate $\mu_{N,h}$ in the $H_0^1(D)$-norm as

$$\|\mathbf{E}[u] - \mu_{N,h}\|_{H_0^1(D)} \leqslant \underbrace{\|\mathbf{E}[u] - \mathbf{E}[u_h]\|_{H_0^1(D)}}_{\text{discretization error}} + \underbrace{\|\mathbf{E}[u_h] - \mu_{N,h}\|_{H_0^1(D)}}_{\text{MC error}}.$$

For the discretization error we obtain, using Jensen's inequality noting that norms are convex function,

$$\|\mathbf{E}[u - u_h]\|_{H_0^1(D)} \leqslant \mathbf{E}\left[\|u - u_h\|_{H_0^1(D)}\right] \leqslant \left(\mathbf{E}\left[\|u - u_h\|_{H_0^1(D)}\right]^2\right)^{1/2}$$

and again for the convex function $\phi(x) = x^2$ to obtain

$$\|\mathbf{E}[u - u_h]\|_{H_0^1(D)} \leqslant \mathbf{E}\left[\|u - u_h\|_{H_0^1(D)}^2\right]^{1/2} = \|u - u_h\|_{L^2(\Omega;H_0^1(D))},$$

which is $O(h)$ by Theorem 2.8.

Monte Carlo Finite Element Method

Error analysis

Theorem 2.9

Under the conditions of Theorem 2.6 there holds

$$\mathbf{E}\left[\|\mathbf{E}[u_h] - \mu_{N,h}\|^2_{H^1_0(D)}\right] \leqslant \frac{K}{N}$$

with a constant K independent of h.

Corollary 2.10

Under the conditions of Theorem 2.6 there holds for any $\epsilon > 0$

$$\mathbf{P}\left(\|\mathbf{E}[u_h] - \mu_{N,h}\|_{H^1_0(D)} \geqslant N^{-1/2+\epsilon}\right) \leqslant LN^{-2\epsilon}$$

for a constant $L > 0$ independent of h.

- **Result:** The total error of estimating the mean $\mathbf{E}[u]$ of the solution of (2.8) using a piecewise linear FE discretization with mesh size h and a MC sample size of N decays at the rate

$$\|\mathbf{E}[u] - \mu_{N,h}\|_{H_0^1(D)} = O(h) + O(N^{-1/2}), \qquad h \to 0,\ N \to \infty.$$

- This is already very slow convergence, and, particularly for low-regularity solutions as arise, e.g., in groundwater flow applications, more advanced techniques such as MLMC methods are attractive.
- Recalling Theorem 2.3, we note that for rough problems we are typically in the regime $\beta < \gamma$. For standard MC estimators on each level ($\delta = 1$) and, as is typical, $\beta = 2\alpha$, we obtain a cost on the order of $\epsilon^{-\gamma/\alpha}$, which is asymptotically the cost of computing one sample on a mesh sufficiently fine to approximate one realization with sufficient spatial accuracy.

Contents

1. Introduction
2. Monte Carlo Methods
3. Random Fields
4. Stochastic Collocation

Contents

Random Fields

- Similar to a stochastic process, a random field is a family of random variables indexed by a parameter. The former concept is often tied to a parameter set which is totally ordered (e.g. $\mathbb{N}$ or $\mathbb{R}_0^+$), whereas for random fields the parameter is a spatial coordinate, typically from subsets of $\mathbb{R}^2$ or $\mathbb{R}^3$.
- Random fields first arose in the field of geostatistics to model phenomena in Earth Sciences such as hydrology, agriculture or geology.
- Since the data for PDE models often consists of one or more functions of space, it is natural to specify the uncertain or random data for PDEs as random fields.
- The alternative view of random fields is as random variables with values in abstract sets, such as spaces of functions, equivalence classes of functions or distributions.
- Naturally, there are extensions to spatio-temporal random fields featuring an additional (ordered) parameter used to model, e.g., turbulence or meteorological phenomena.

Definition 3.1

Given a set $D \subset \mathbb{R}^d$, $d \in \mathbb{N}$ and a probability space $(\Omega, \mathfrak{A}, \mathbf{P})$, a (real-valued) random field is a mapping

$$a : D \times \Omega \to \mathbb{R}$$

such that each function $a(\boldsymbol{x}, \cdot) : \Omega \to \mathbb{R}$, $\boldsymbol{x} \in D$, is a random variable.

Random Fields

Random field as a function-valued random variable

Definition 3.2

For each fixed $\omega \in \Omega$ the associated function $a(\cdot, \omega) : D \to \mathbb{R}$ is called a realization of the random field.

- Denote by $\mathbb{R}^D$ the set of all real-valued functions $f : D \to \mathbb{R}$. In particular, realizations of a real-valued random field belong to $\mathbb{R}^D$ by Definition 3.2.
- Denote further by $\mathfrak{A}(\mathbb{R}^D)$ the smallest σ-algebra containing all sets

$$A = \{f \in \mathbb{R}^D : (f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n)) \in B\}$$

for any $B \in \mathfrak{B}(\mathbb{R}^n), \boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in D, n \in \mathbb{N}$.

Proposition 3.3

Let a be a random field on $D \subset \mathbb{R}^d$ with underlying probability space $(\Omega, \mathfrak{A}, \mathbf{P})$. Then the mapping $\omega \mapsto a(\cdot, \omega)$ from $(\Omega, \mathfrak{A})$ to the measurable space $(\mathbb{R}^D, \mathfrak{A}(\mathbb{R}^D))$ is measurable and hence a random variable with values in $\mathbb{R}^D$.

Random Fields

Independence

Definition 3.4

(a) Two real-valued random fields $\{a(\boldsymbol{x}), \boldsymbol{x} \in D\}$ and $\{b(\boldsymbol{y}), \boldsymbol{y} \in D\}$ on $D \subset \mathbb{R}^d$ are said to be independent if the associated $(\mathbb{R}^D, \mathfrak{A}(\mathbb{R}^D))$-valued random variables are independent.

(b) We call $f_i : D \to \mathbb{R}$, $i = 1, 2$, independent realizations of a real-valued random field a on $D \subset \mathbb{R}^d$ if $f_i(\boldsymbol{x}) = a_i(\boldsymbol{x}, \omega)$ for some $\omega \in \Omega$, where a_i are i.i.d. random fields with the same distribution as a.

Random Fields

Finite-dimensional distributions

Definition 3.5

For a real-valued random field $\{a(\boldsymbol{x}), \boldsymbol{x} \in D\}$ defined on $D \subset \mathbb{R}^d$ the probability distributions of all random vectors $(a(\boldsymbol{x}_1), \dots, a(\boldsymbol{x}_n))$ with $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in D$ on $(\mathbb{R}^n, \mathfrak{B}(\mathbb{R}^n))$ are known as the finite-dimensional distributions of a.

The Daniell-Kolmogorov theorem states consistency conditions for defining the distribution of a random field by a family of finite-dimensional probability measures. We denote by $\mathbf{P}_a$ the probability distribution of a random field a on the measurable space $(\mathbb{R}^D, \mathfrak{A}(\mathbb{R}^D))$.

Random Fields

Characterization by finite-dimensional distributions

Theorem 3.6 (Daniell & Kolmogorov)

Suppose that for each set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \subset D$ there exists a probability measure $\mu_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n}$ on $\mathbb{R}^n$ such that

(i) For any permutation σ of $\{1, \dots, n\}$ and any $B \in \mathfrak{B}(\mathbb{R}^n)$ there holds

$$\mu_{\boldsymbol{x}_{\sigma(1)},\dots,\boldsymbol{x}_{\sigma(n)}}(\sigma(B)) = \mu_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n}(B),$$

where $\sigma(B) = \{(\boldsymbol{x}_{\sigma(1)}, \dots, \boldsymbol{x}_{\sigma(n)}) : (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \in B\}$ and

(ii) for $m < n$ and any $B \in \mathfrak{B}(\mathbb{R}^m)$

$$\mu_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n}(B \times \mathbb{R}^{n-m}) = \mu_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_m}(B).$$

Then there exists a random field $\{a(\boldsymbol{x}), \boldsymbol{x} \in D\}$ with finite-dimensional distributions $\mu_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n}$. If $a(\boldsymbol{x})$ and $b(\boldsymbol{x})$ are two such random fields, then $\mathbf{P}_a(A) = \mathbf{P}_b(A)$ for any $A \in \mathfrak{A}(\mathbb{R}^D)$.

Random Fields

Mean, covariance

Definition 3.7

A random field a on $D \subset \mathbb{R}^d$ is said to be of second order if for all $\boldsymbol{x} \in D$ there holds $a(\boldsymbol{x}) = a(\boldsymbol{x}, \cdot) \in L^2(\Omega; \mathbb{R})$. We say a second-order random field a has mean function $\overline{a}(\boldsymbol{x}) := \mathbf{E}\left[a(\boldsymbol{x})\right]$ and covariance function

$$c(\boldsymbol{x}, \boldsymbol{y}) = c_a(\boldsymbol{x}, \boldsymbol{y}) := \mathbf{Cov}(a(\boldsymbol{x}), a(\boldsymbol{y})), \qquad \boldsymbol{x}, \boldsymbol{y} \in D.$$

Definition 3.8

A function $f : D \times D \to \mathbb{R}$ is called positive semidefinite if for any n-tuple $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \in D^n$ and vector $\boldsymbol{z} = [z_1, \dots, z_n]^\mathsf{T} \in \mathbb{R}^n$ there holds

$$\sum_{j,k=1}^{n} z_j z_k f(\boldsymbol{x}_j, \boldsymbol{x}_k) \geqslant 0.$$

Note: In the stochastics literature this property is often called simply *positive definite*.

Random Fields

Covariance functions and positive definiteness

Theorem 3.9

Let $D \subset \mathbb{R}^d$. The following statements are equivalent:

(a) There exists a real-valued second-order random field $\{a(\boldsymbol{x}), \boldsymbol{x} \in D\}$ with covariance function $c : D \times D \to \mathbb{R}$.

(b) $c \in \mathbb{R}^{D \times D}$ is symmetric and positive semidefinite.

Definition 3.10

A real-valued random field on $D \subset \mathbb{R}^d$ is called Gaussian if each random vector $[a(\boldsymbol{x}_1), \dots, a(\boldsymbol{x}_n)]$ follows an n-variate normal distribution for any $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in D$ and any $n \in \mathbb{N}$.

Corollary 3.11

The probability distribution $\mathbf{P}_a$ on $(\mathbb{R}^D, \mathfrak{A}(\mathbb{R}^D))$ of a real-valued Gaussian random field a is uniquely determined by its mean and covariance function.

Contents

1. Introduction

2. Monte Carlo Methods

3. Random Fields

   3.1 Introduction

   3.2 Karhunen-Loève Expansion

   3.3 Regularity of Random Fields

   3.4 Covariance Eigenvalue Decay

4. Stochastic Collocation

Random Fields

Values in $L^2(D)$

Given a second-order random field a on $D \subset \mathbb{R}^d$ with mean $\overline{a}$, consider the centered random field $a - \overline{a}$. Given a CONS $\{\psi_m\}_{m\in\mathbb{N}}$ of $L^2(D)$, we have for each realization of a:

$$a(\cdot,\omega) - \overline{a} = \sum_{m=1}^{\infty} \xi_m(\omega)\psi_m,$$

where the ξ_m are random variables defined by

$$\xi_m(\omega) := (a(\cdot,\omega) - \overline{a}, \psi_m)_{L^2(D)}.$$

The Karhunen-Loève expansion of a results from choosing as a particular CONS the eigenfunctions of the covariance operator $C = C_a : L^2(D) \to L^2(D)$ of a, which is given by

$$u \mapsto Cu, \qquad (Cu)(\boldsymbol{x}) = \int_D u(\boldsymbol{y})c(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}, \quad \boldsymbol{x} \in D. \tag{3.1}$$

Lemma 3.12

If $a \in L^2(\Omega; L^2(D))$, then $\overline{a} \in L^2(D)$ and $a(\cdot,\omega) \in L^2(D)$ **P**-a.s.

By Definition A.23, if $a \in L^2(\Omega; L^2(D))$ for $D \subset \mathbb{R}^d$, we have for any $\phi, \psi \in L^2(D)$ by Fubini's theorem

$$\begin{aligned}(C\phi, \psi)_{L^2(D)} &= \mathbf{Cov}\big((\phi, a)_{L^2(D)}, (\psi, a)_{L^2(D)}\big) \\ &= \mathbf{E}\left[\int_D \phi(\boldsymbol{x})[a(\boldsymbol{x}) - \overline{a}(\boldsymbol{x})]\,\mathrm{d}\boldsymbol{x} \int_D \psi(\boldsymbol{y})[a(\boldsymbol{y}) - \overline{a}(\boldsymbol{y})]\,\mathrm{d}\boldsymbol{y}\right] \\ &= \int_D \int_D \mathbf{Cov}(a(\boldsymbol{x}), a(\boldsymbol{y}))\,\phi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\;\psi(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y},\end{aligned}$$

from which we infer that the covariance operator C_a of the $L^2(D)$-valued random variable a is the linear integral operator $C_a : L^2(D) \to L^2(D)$ with kernel function $c_a \in L^2(D \times D)$ given by $c_a(\boldsymbol{x}, \boldsymbol{y}) = \mathbf{Cov}(a(\boldsymbol{x}), a(\boldsymbol{y}))$ (not pointwise, in general !)

Example 3.13

For $d = 1$ and $D = [-b, b]$, $b > 0$, the exponential covariance function is defined by

$$c(x, y) = e^{\frac{-|x-y|}{\ell}}, \qquad \ell > 0.$$

The eigenvalues of the associated covariance operator are given by

$$\lambda_m = \frac{2\ell}{\ell^2 \omega_m^2 + 1}, \ (m \text{ even}), \qquad \lambda_m = \frac{2\ell}{\ell^2 \tilde{\omega}_m^2 + 1}, \ (m \text{ odd})$$

where ω_m and $\tilde{\omega}_m$ denote the solutions of the transcendental equations

$$1 - \omega\ell \tan(\omega b) = 0 \quad \text{and} \quad \omega\ell + \tan(\omega b) = 0,$$

respectively. The associated eigenfunctions are given by

$$f_m(x) = \frac{\cos(\omega_m x)}{\sqrt{b + \frac{\sin(2\omega_m b)}{2\omega_m}}}, \qquad \tilde{f}_m(x) = \frac{\sin(\tilde{\omega}_m x)}{\sqrt{b - \frac{\sin(2\tilde{\omega}_m b)}{2\tilde{\omega}_m}}},$$

Theorem 3.14

If $(\lambda_m, a_m)_{m\in\mathbb{N}}$ denotes the sequence of eigenpairs (in descending order, $\|a_m\|_{L^2(D)} = 1$) of the covariance operator C_a associated with the random field $a \in L^2(\Omega; L^2(D))$ with mean function $\overline{a}(\boldsymbol{x})$, then

$$a(\boldsymbol{x}, \omega) = \overline{a}(\boldsymbol{x}) + \sum_{m=1}^{\infty} \sqrt{\lambda_m}\, a_m(\boldsymbol{x})\, \xi_m(\omega), \tag{3.2}$$

where the series converges in $L^2(\Omega; L^2(D))$, the random variables

$$\xi_m(\omega) = \frac{1}{\sqrt{\lambda_m}} (a(\cdot, \omega) - \overline{a}, a_m)_{L^2(D)}$$

have mean zero, unit variance and are pairwise uncorrelated.
If the random field is, in addition, Gaussian, then $\xi_m \sim \mathrm{N}(0, 1)$ are i.i.d.

- The KL expansion suggests a convenient approach for approximating a random field to a specified accuracy by truncation:

$$a(\boldsymbol{x},\omega) \approx a_M(\boldsymbol{x},\omega) := \overline{a}(\boldsymbol{x}) + \sum_{m=1}^{M} \sqrt{\lambda_m}\, a_m(\boldsymbol{x})\, \xi_m(\omega). \tag{3.3}$$

- The truncated RF a_M has the same mean as a and the covariance function

$$c(\boldsymbol{x},\boldsymbol{y}) \approx c_M(\boldsymbol{x},\boldsymbol{y}) = c_{a,M}(\boldsymbol{x},\boldsymbol{y}) = \sum_{m=1}^{M} \lambda_m a_m(\boldsymbol{x}) a_m(\boldsymbol{y}), \qquad \boldsymbol{x},\boldsymbol{y} \in D. \tag{3.4}$$

- Since $\{\phi_j(\boldsymbol{x})\phi_k(\boldsymbol{u})\}_{j,k\in\mathbb{N}}$ is a CONS for $L^2(D\times D)$, $c_M \to c$ in $L^2(D\times D)$.
- For a bounded domain D and continuous covariance function, its series expansion (3.4) and the MSE of the truncated KL series (3.3) both converge uniformly.

Random Fields

KL expansion, uniform convergence

Theorem 3.15

Let a_M denote the truncated approximation defined in (3.3) of a real-valued random field $a \in L^2(\Omega; L^2(D))$ for a compact domain $D \subset \mathbb{R}^d$ with covariance function $c \in C(D \times D)$. Then the KL eigenfunctions $\{a_m\}_{m\in\mathbb{N}}$ are continuous on D and the series expansion of c converges uniformly, i.e.,

$$\sup_{\boldsymbol{x},\boldsymbol{y}\in D} |c(\boldsymbol{x},\boldsymbol{y}) - c_M(\boldsymbol{x},\boldsymbol{y})| \leqslant \sup_{\boldsymbol{x}\in D} \sum_{m=M+1}^{\infty} \lambda_m a_m(\boldsymbol{x})^2 \to 0, \qquad M \to \infty. \quad (3.5)$$

In addition,

$$\sup_{\boldsymbol{x}\in D} \mathbf{E}\left[(a(\boldsymbol{x}) - a_M(\boldsymbol{x}))^2\right] \to 0 \qquad M \to \infty.$$

- For the variance of the truncated KL expansion, we have

$$\mathbf{Var}(a(\boldsymbol{x})) - \mathbf{Var}(a_M(\boldsymbol{x})) = \mathbf{E}\left[(a(\boldsymbol{x}) - a_m(\boldsymbol{x}))^2\right] = \sum_{m=M+1}^{\infty} \lambda_m a_m(\boldsymbol{x})^2 \geqslant 0,$$

  hence a_M always underestimates the variance of a.
- Viewed as a random variable $a \in L^2(\Omega; L^2(D))$, we have for the truncation error

$$\|a - a_M\|^2_{L^2(\Omega;L^2(D))} = \sum_{m=M+1}^{\infty} \lambda_m.$$

- In addition,

$$\begin{aligned}\|a - a_M\|^2_{L^2(\Omega;L^2(D))} &= \mathbf{E}\left[\|a - \overline{a}\|^2_{L^2(D)} - \|a_M - \overline{a}\|^2_{L^2(D)}\right] \\ &= \int_D \mathbf{Var}\, a(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \sum_{m=1}^{M} \lambda_m.\end{aligned}$$

Random Fields

KL expansion, variance

- This allows an assessment of the truncation error w.r.t. $\|\cdot\|_{L^2(\Omega;L^2(D))}$ provided the first M eigenvalues can be calculated as well as the integral of **Var** a over D.
- Eigenvalue approximations can be obtained by solving the covariance eigenproblem numerically. If **Var** $a \equiv \sigma^2$ on D, this yields

$$\|a - a_M\|^2_{L^2(\Omega;L^2(D))} = \sigma^2|D| - \sum_{m=1}^{M} \lambda_m,$$

where $|D| = \int_D \mathrm{d}\boldsymbol{x}$ is the Lebesgue measure of D.
In this case the error can always be estimated once a number of leading eigenvalues are available.

Contents

1. Introduction

2. Monte Carlo Methods

3. Random Fields

   3.1 Introduction

   3.2 Karhunen-Loève Expansion

   3.3 Regularity of Random Fields

   3.4 Covariance Eigenvalue Decay

4. Stochastic Collocation

Regularity of Random Fields

Definition 3.16

A random field $\{a(\boldsymbol{x}), \boldsymbol{x} \in D\}$ is said to be mean-square continuous if, for all $\boldsymbol{x} \in D$,

$$\|a(\boldsymbol{x}+\boldsymbol{h}) - a(\boldsymbol{x})\|_{L^2(\Omega)} = \mathbf{E}\left[(a(\boldsymbol{x}+\boldsymbol{h}) - a(\boldsymbol{x}))^2\right] \to 0 \quad \text{as } \boldsymbol{h} \to \mathbf{0}.$$

We assume centered random fields a in the remainder of this section, i.e., $\overline{a} \equiv 0$.

Theorem 3.17

Let $\{a(\boldsymbol{x}), \boldsymbol{x} \in D\}$ be a centered random field. Then its covariance function c is continuous at $(\boldsymbol{x}, \boldsymbol{x})$, $\boldsymbol{x} \in D$, if and only if $\mathbf{E}\left[(a(\boldsymbol{x}+\boldsymbol{h}) - a(\boldsymbol{x}))^2\right] \to 0$ as $\boldsymbol{h} \to \mathbf{0}$. In particular, if $c \in C(D \times D)$, then a is mean-square continuous.

Corollary 3.18

Let $\{a(\boldsymbol{x}), \boldsymbol{x} \in D\}$ be a centered random field. If its covariance function is continuous along the 'diagonal' $\{(\boldsymbol{x}, \boldsymbol{x}) : \boldsymbol{x} \in D\}$, then it is continuous throughout $D \times D$.

Regularity of Random Fields

Mean-square differentiability

Theorem 3.19

Let $\{a(\boldsymbol{x}), \boldsymbol{x} \in D\}$ be a centered second-order random field. If its covariance function $c \in C^2(D \times D)$, then a is mean-square differentiable, i.e., there exists a random field $\{\partial_{x_j} a(\boldsymbol{x}), \boldsymbol{x} \in D\}$ such that for all $j = 1, 2, \dots, d$,

$$\left\| \frac{a(\boldsymbol{x} + h\boldsymbol{e}_j) - a(\boldsymbol{x})}{h} - \partial_{x_j} a(\boldsymbol{x}) \right\|_{L^2(\Omega)} \to 0 \quad \text{as } h \to 0$$

and $\partial_{x_j} a(\boldsymbol{x})$ has covariance function

$$c_j(\boldsymbol{x}, \boldsymbol{y}) = \frac{\partial^2 c(\boldsymbol{x}, \boldsymbol{y})}{\partial x_j \partial y_j}.$$

Analogous relations hold for higher order mean-square derivatives of a random field given higher order differentiability of the covariance function.

Regularity of Random Fields

Regularity of realizations

- Mean-square continuity and differentiability depend on the expectation, i.e., on an average over all realizations.
- A related issue is the regularity of each individual realization.
- Even for Gaussian fields, it is not possible to show that each realization is continuous.
- Even though the distribution of a Gaussian random field is uniquely defined on $\mathfrak{A}(\mathbb{R}^D)$, realization-wise continuity cannot hold in general.
- $\mathfrak{A}(\mathbb{R}^D)$ is contructed from a countable set of conditions, whereas statements about the continuity of functions involve conditions on a continuum of points, i.e., uncountably many conditions.
- However, given a condition of the moments of the 'increments' $a(\boldsymbol{x}) - a(\boldsymbol{y})$, a version of $a(\boldsymbol{x})$ with continuous realizations can be shown to exist.

Regularity of Random Fields

Regularity of realizations

Theorem 3.20

Let D be a bounded domain in $\mathbb{R}^d$ and $\{a(\boldsymbol{x}), \boldsymbol{x} \in D\}$ be a centered Gaussian random field such that, for some $L, s > 0$

$$\mathbf{E}\left[|a(x) - a(y)|^2\right] \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\|_2^s \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \overline{D}.$$

Then for any $p \geqslant 1$ there exists a random variable K such that $e^K \in L^p(\Omega)$ and

$$|a(\boldsymbol{x}) - a(\boldsymbol{y})| \leqslant K(\omega)\|\boldsymbol{x} - \boldsymbol{y}\|_2^{(s-\epsilon)/2} \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \overline{D} \text{ a.s.},$$

i.e., realizations of a are Hölder continuous with exponent $s/2$.

Regularity of Random Fields

Regularity of realizations

Definition 3.21

A random field $\{b(\boldsymbol{x}), \boldsymbol{x} \in D\}$ is called a version of a random field $\{a(\boldsymbol{x}), \boldsymbol{x} \in D\}$ if

$$\mathbf{P}\left(a(\boldsymbol{x}) = b(\boldsymbol{x})\right) = 1 \qquad \forall \boldsymbol{x} \in D.$$

A random field $\{a(\boldsymbol{x}), \boldsymbol{x} \in D\}$ is said to have a continuous version if there exists a version of a with continuous realizations.

Theorem 3.22 (cf. [Kallenberg (1997)], Thm. 2.23)

Let a be a random field on $D \subset \mathbb{R}^d$ with values in a Banach space and assume for some $a, b > 0$ that

$$\mathbf{E}\left[\|a(\boldsymbol{x}) - a(\boldsymbol{y})\|^a\right] \lesssim \|\boldsymbol{x} - \boldsymbol{y}\|^{d+b}, \qquad \boldsymbol{x}, \boldsymbol{y} \in D.$$

Then a has a continuous version, and for any $c \in (0, b/a)$ the latter is a.s. locally Hölder continuous with exponent c.

Contents

1. Introduction

2. Monte Carlo Methods

3. Random Fields

   3.1 Introduction

   3.2 Karhunen-Loève Expansion

   3.3 Regularity of Random Fields

   3.4 Covariance Eigenvalue Decay

4. Stochastic Collocation

Covariance Eigenvalue Decay

Stationarity

Definition 3.23

(a) A random field a is strictly stationary or homogeneous if its finite-dimensional distributions are invariant under translation, i.e., if the multivariate distribution of $(a(\boldsymbol{x}_1), \dots, a(\boldsymbol{x}_n))$ is the same as that of $(a(\boldsymbol{x}_1 + \boldsymbol{h}), \dots, a(\boldsymbol{x}_n + \boldsymbol{h}))$, for all $\boldsymbol{h}$.

(b) A random field a is (wide-sense) stationary or (wide-sense) homogeneous if its mean is constant and its covariance function satisfies $c(\boldsymbol{x}, \boldsymbol{y}) = c(\boldsymbol{x} - \boldsymbol{y})$. Such a covariance function is known as a stationary covariance.

Example 3.24

The separable exponential covariance function is given by

$$c(\boldsymbol{x}, \boldsymbol{y}) = \prod_{j=1}^{d} e^{\frac{-|x_j - y_j|}{\ell_j}}, \qquad \ell_j > 0,$$

where ℓ_j is a correlation length parameter in the j-th Cartesian direction, is an example of a stationary covariance function.

Covariance Eigenvalue Decay

Fourier representation

Theorem 3.25 (Wiener-Khintchine)

The following two statements are equivalent:

(a) There exists a mean-square continuous stationary random field $\{a(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d\}$ with stationary covariance function c.

(b) The function $c : \mathbb{R}^d \to \mathbb{R}$ is such that

$$c(\boldsymbol{x}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\lambda}\cdot\boldsymbol{x}}\,\mathrm{d}F(\boldsymbol{\lambda})$$

for some measure F on $\mathbb{R}^d$ with $F(\mathbb{R}^d) < \infty$,

The measure F is called the spectral distribution. If it exists, the density f of F is called the spectral density. Alternatively, given $c : \mathbb{R}^d \to \mathbb{R}$, we may compute

$$f(\boldsymbol{\lambda}) = (2\pi)^{-d} \int_{\mathbb{R}^d} e^{-i\boldsymbol{\lambda}\cdot\boldsymbol{x}} c(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

If f is nonnegative and integrable then c is a valid covariance function.

Covariance Eigenvalue Decay

Fourier representation

Example 3.26 (Separable exponential covariance)

The Fourier transform of the separable covariance function is obtained as the rpoduct of the transforms of its factors, i.e.,

$$f(\boldsymbol{\lambda}) = (2\pi)^{-d} \int_{\mathbb{R}^d} e^{-i\boldsymbol{\lambda}\cdot\boldsymbol{x}} c(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \prod_{j=1}^{d} (2\pi)^{-1} \int_{\mathbb{R}} e^{-ix_j\lambda_j} e^{-|x_j|/\ell_j} \, \mathrm{d}x_j,$$

yielding

$$f(\boldsymbol{\lambda}) = \prod_{j=1}^{d} \frac{\ell_j}{\pi(\lambda_j^2 + \ell_j^2)}.$$

Since $\ell_j > 0$ for all j, f is nonnegative and is the density of a measure F with $F(\mathbb{R}^d) < \infty$. By the Wiener-Khintchine theorem c is thus the covariance kernel for some mean-square continuous random field.

Covariance Eigenvalue Decay

Fourier representation

Example 3.27 (Gaussian covariance)

For a symmetric positive definite matrix $A \in \mathbb{R}^{d\times d}$ the function

$$c(\boldsymbol{x}) = e^{-\boldsymbol{x}^\mathsf{T} A \boldsymbol{x}}, \qquad \boldsymbol{x} \in \mathbb{R}^d,$$

has the Fourier transform

$$f(\boldsymbol{\lambda}) = (2\pi)^{-d} \int_{\mathbb{R}^d} e^{-i\boldsymbol{x}\cdot\boldsymbol{\lambda}} e^{-\boldsymbol{x}^\mathsf{T} A \boldsymbol{x}} \, \mathrm{d}\boldsymbol{x} = \frac{1}{(2\pi)^{d/2} 2^{d/2} \sqrt{\det A}} = e^{-\boldsymbol{\lambda}^\mathsf{T} A^{-1} \boldsymbol{\lambda}/4}.$$

f is nonnegative and is the density of a measure F – in fact the Gaussian distribution $\mathrm{N}(\mathbf{0}, 2A)$. Again, the Wiener-Khintchine theorem asserts that c is the covariance function of a random field.

Definition 3.28

A stationary random field $\{a(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d\}$ is said to be isotropic if its covariance function is invariant under rotations, i.e.,

$$c(\boldsymbol{x}, \boldsymbol{y}) = c(r), \qquad r = \|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

Example 3.29 (Isotropic Gaussian covariance)

A simple example of an isotropic covariance function is $c(r) = e^{-r^2}$, arising from a Gaussian covariance with $A = I_d$. In fact, $c(\boldsymbol{x}) = e^{-\boldsymbol{x}^\top A \boldsymbol{x}}$ is isotropic whenever $A = \sigma I_d$ for some $\sigma > 0$.

Example 3.30 (Bessel covariance)

Another isotropic covariance function, proposed by Whittle as a generalization of the exponential covariance to higher dimensions, is given by $c(r) = rK_1(r)$, where K_1 is the modified Bessel function of second kind with index 1.

For isotropic functions the Fourier transform in the Wiener-Khintchine theorem becomes a Hankel transform (cf. Theorem D.1). For $f(s) = f(\|\boldsymbol{\lambda}\|_2)$, we obtain for $d = 1, 2, 3$

$$c(r) = \begin{cases} 2\int_0^\infty \cos(rs)\, f(s)\,\mathrm{d}s, & d = 1, \\[2ex] 2\pi\int_0^\infty J_0(rs) f(s) s\,\mathrm{d}s, & d = 2, \\[2ex] 4\pi\int_0^\infty \frac{1}{rs}\sin(rs) f(s) s^2\,\mathrm{d}s, & d = 3. \end{cases}$$

Covariance Eigenvalue Decay

Isotropy

Theorem 3.31

Let $\{a(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d\}$ be an isotropic random field with mean-square continuous covariance function c. There exists a finite measure F on $\mathbb{R}^+$ known as the radial spectral distribution such that

$$c(r) = \Gamma\left(\tfrac{d}{2}\right) \int_0^\infty \frac{J_\nu(rs)}{(\frac{rs}{2})^\nu} \,\mathrm{d}F(s), \qquad \nu = \tfrac{d}{2} - 1.$$

If the spectral density exists, $f(s) = f(\boldsymbol{\lambda})$ for $s = \|\boldsymbol{\lambda}\|_2$ is called the radial spectral density function. Then

$$\mathrm{d}F(s) = \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})} s^{d-1} f(s)\,\mathrm{d}s \quad \text{and} \quad f(s) = (2\pi)^{-d/2} \int_0^\infty \frac{J_\nu(rs)}{(rs)^\nu} c(r) r^{d-1} \,\mathrm{d}r.$$

Covariance Eigenvalue Decay

The Matérn class

The Matérn class is a family of isotropic covariance functions named after the Swedish forestry statistician Bertil Matérn and is very popular in geostatistics as well as machine learning etc.

The covariance function is given by

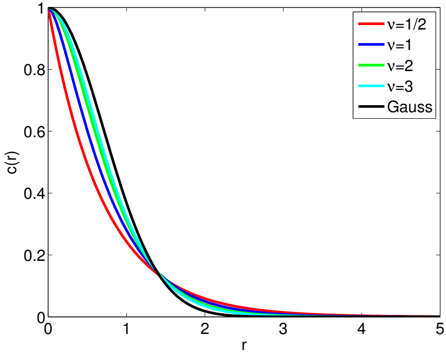
$$c(r) = \frac{\sigma^2}{2^{\nu-1}\,\Gamma(\nu)} \left(\frac{2\sqrt{\nu}\,r}{\rho}\right)^{\nu} K_\nu\left(\frac{2\sqrt{\nu}\,r}{\rho}\right), \qquad r = \|\boldsymbol{x} - \boldsymbol{y}\|_2, \tag{3.6}$$

where

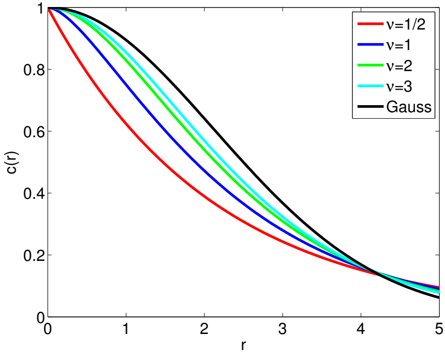
K_ν	is the modified (second-kind) Bessel function of order ν,
ν	is known as the smoothness parameter,
σ^2	is the variance parameter,
ρ	is the correlation length parameter,
Γ	denotes the Gamma-function.

Covariance Eigenvalue Decay

The Matérn class



$\rho = 1$



$\rho = 3$

- With smaller correlation length ρ the Matérn covariance function becomes more strongly concentrated near $r = 0$.
- With increasing values of the smoothness parameter ν the Matérn covariance function becomes smoother at $r = 0$. (It is analytic everywhere else.)

Covariance Eigenvalue Decay

Decay rate

The Matérn family has a number of attractive features:

- It contains the exponential, Bessel and Gaussian covariance functions as special cases:

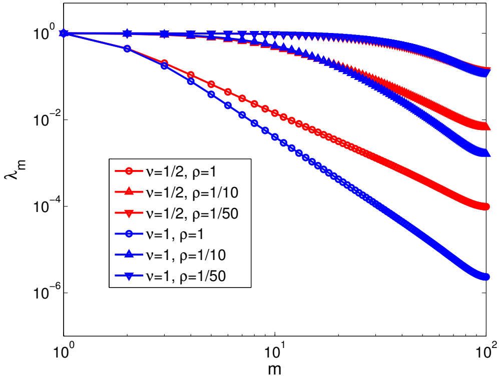
$$
\begin{array}{lll}
\nu = \frac{1}{2}: & c(r) = \sigma^2 \exp(-\sqrt{2}r/\rho) & \text{exponential covariance} \\
\nu = 1: & c(r) = \sigma^2 \left(\frac{2r}{\rho}\right) K_1 \left(\frac{2r}{\rho}\right) & \text{Bessel covariance} \\
\nu \to \infty: & c(r) = \sigma^2 \exp(-r^2/\rho^2) & \text{Gaussian covariance}
\end{array}
$$

- Smoothness of realizations: a random field with Matérn covariance function is s times mean-square differentiable if and only if $\nu > s$.
- The flexibility of the parametrization allows its application to many statistical situation, the parameters may be estimated from observed data using statistical techniques.

Covariance Eigenvalue Decay

Matérn eigenvalue decay

Before asymptotic decay sets in (determined by the smoothness of the kernel), there is a preasymptotic plateau whose length is determined by the correlation length parameter ρ.



Eigenvalue decay, Matérn covariance kernel, $D = [-1, 1]$.

Covariance Eigenvalue Decay

Widom's result

In a paper published in 1963[2], Harold Widom analyzed linear integral operators of the form

$$u \mapsto Ku, \qquad (Ku)(\boldsymbol{x}) = \int_{\mathbb{R}^d} V(\boldsymbol{x})^{1/2} k(\boldsymbol{x}-\boldsymbol{y}) V(\boldsymbol{y})^{1/2} u(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}. \tag{3.7}$$

We obtain the covariance operator for an isotropic covariance function on a bounded domain $D \subset \mathbb{R}^d$ by setting $V(\boldsymbol{x}) = \mathbb{1}_D$ and $k(\boldsymbol{x}-\boldsymbol{y}) = c(\|\boldsymbol{x}-\boldsymbol{y}\|_2)$.

Definition 3.32

Two functions $f : E \to \mathbb{R}$ and $g : F \to \mathbb{R}$, $E \subset \mathbb{R}^n$ and $F \subset \mathbb{R}^m$ are said to be equimeasurable if, for all $t \in \mathbb{R}$,

$$|\{\boldsymbol{x} \in E : f(\boldsymbol{x}) > t\}| = |\{\boldsymbol{y} \in F : g(\boldsymbol{y}) > t\}|$$

where $|\cdot|$ denotes Lebesgue measure.

[2] Widom, H., Asymptotic behavior of the eigenvalues of certain integral equations. *Trans. Amer. Math. Soc.* 109, 278–295 (1963).

Covariance Eigenvalue Decay

Widom's result

We denote the spectral density (Fourier transform) of $c = c(\boldsymbol{x}) = c(\|\boldsymbol{x}\|_2)$ by

$$\hat{c}(\boldsymbol{\lambda}) = (2\pi)^{-d} \int_{\mathbb{R}^d} c(\boldsymbol{x}) e^{-i\boldsymbol{x}\cdot\boldsymbol{\lambda}} \, \mathrm{d}\boldsymbol{x} = f(s)$$

and set $K(\boldsymbol{\lambda}) := (2\pi)^d \, \hat{c}(\boldsymbol{\lambda})$.

Theorem 3.33 (Widom, 1963)

For the integral operator K in (3.7) let V be a bounded, nonnegative function with bounded support, let k be integrable over $\mathbb{R}^d$ with an ultimately positive Fourier transform and let $\{\lambda_m\}_{m\in\mathbb{N}}$ denote its (nonincreasing) sequence of eigenvalues. If the function $\phi_0 : \mathbb{R}_0^+ \to \mathbb{R}$ is equimeasurable to $V(\boldsymbol{x})K(\boldsymbol{\lambda}) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, then

$$\lambda_m \asymp \phi_0((2\pi)^d m) \qquad \text{as } m \to \infty.$$

Note: The theorem still holds when the integral operator leads to a $K'(\boldsymbol{\lambda})$ such that

$$K(\boldsymbol{\lambda}) \asymp K'(\boldsymbol{\lambda}), \qquad \text{as } \|\boldsymbol{\lambda}\|_2 \to \infty.$$

Covariance Eigenvalue Decay

Decay rate

Corollary 3.34

Let $c = c(r)$ be an isotropic covariance function on $\mathbb{R}^d$ with radial spectral density $f = f(s)$. Assume that $f(s) \asymp bs^{-\rho}$ as $s \to \infty$, for some $b, \rho > 0$. Let D be a bounded domain in $\mathbb{R}^d$ and let $\{\lambda_m\}_{m\in\mathbb{N}}$ denote the (nonincreasing) eigenvalues of the covariance operator C given by (3.1). Then

$$\lambda_m \asymp K(D, d, \rho, b)\, m^{-\rho/d}, \qquad m \to \infty,$$

with $K(D, d, \rho, b) := (2\pi)^{d-\rho} b(|D|V_d)^{\rho/d}$, where $V_d = \frac{2\pi^{d/2}}{d\,\Gamma(d/2)}$ denotes the volume of the unit sphere in $\mathbb{R}^d$.

Corollary 3.35

If the spectral density $f(s)$ of an isotropic random field satisfies $f(s) \asymp bs^{-\rho}$, then

$$\lambda_m \asymp K m^{-\rho/d}, \qquad m \to \infty,$$

with $K = (2\pi)^{d-\rho} b\,(|D|V_d)^{\rho/d}$ where V_d denotes the volume of the d-dimensional unit sphere.

Covariance Eigenvalue Decay

Decay rate

The Fourier transform of the Matérn covariance function with smoothness parameter ν, variance σ^2 and correlation length parameter ℓ is given by

$$f(s;\nu,\sigma,\ell) = \sigma^2 \pi^{-d/2} \frac{\Gamma(\nu + d/2)}{\Gamma(\nu)} \frac{\alpha^{2\nu}}{(s^2+\alpha^2)^{\nu+d/2}}.$$

where $\alpha := 2\sqrt{\nu}/\ell$.

Corollary 3.36

For the Matérn covariance in d dimensions with smoothness parameter ν the covariance eigenvalues decay asymptotically like

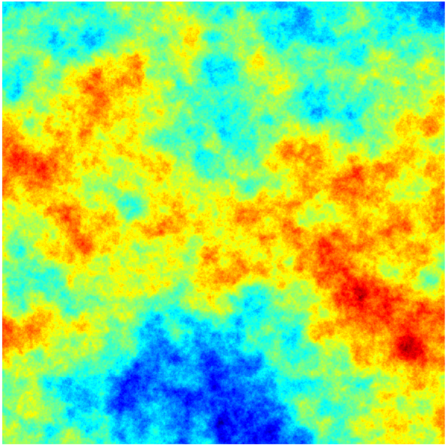
$$\lambda_m \asymp K m^{-(1+2\nu/d)}, \qquad m \to \infty,$$

where

$$K = (2\pi)^{-2\nu-d/2} \sigma^2 \frac{\Gamma(\nu + d/2)}{\Gamma(\nu)} \left(\frac{2\sqrt{\nu}}{\ell}\right)^{2\nu} (|D|V_d)^{1+2\nu/d}.$$

Covariance Eigenvalue Decay

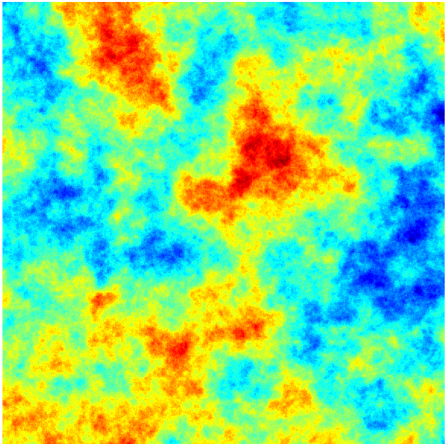
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.5$

Covariance Eigenvalue Decay

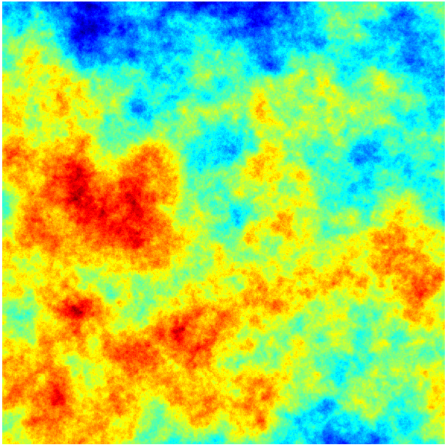
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.5$

Covariance Eigenvalue Decay

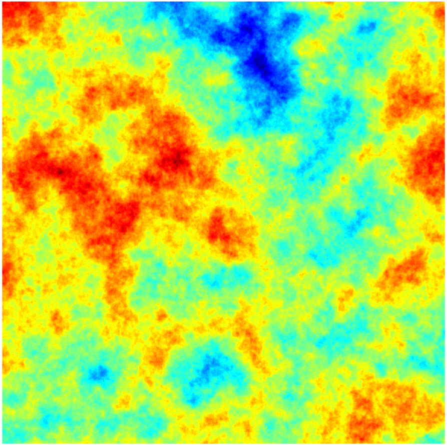
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.5$

Covariance Eigenvalue Decay

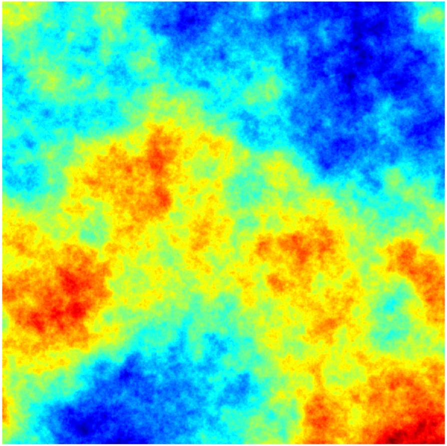
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.5$

Covariance Eigenvalue Decay

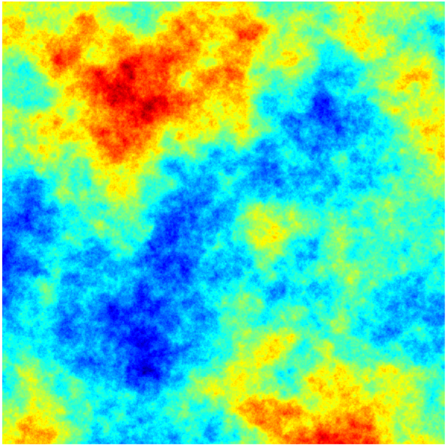
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.5$

Covariance Eigenvalue Decay

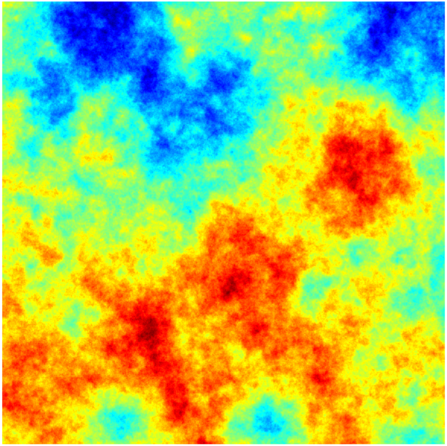
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.5$

Covariance Eigenvalue Decay

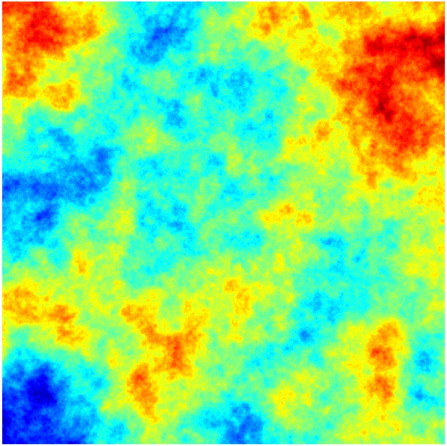
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.5$

Covariance Eigenvalue Decay

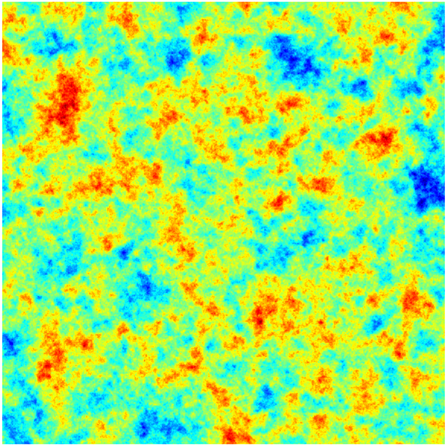
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.5$

Covariance Eigenvalue Decay

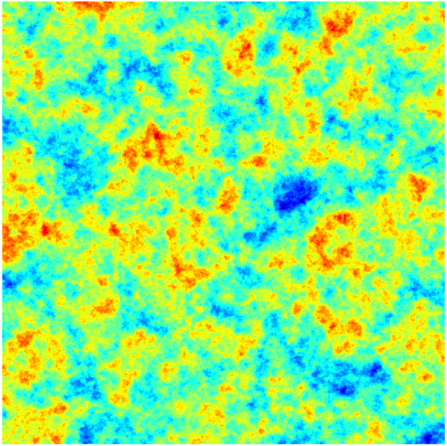
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

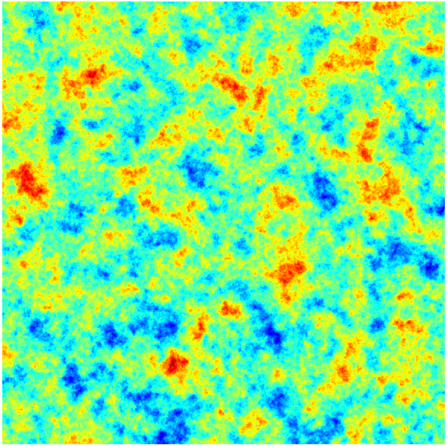
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

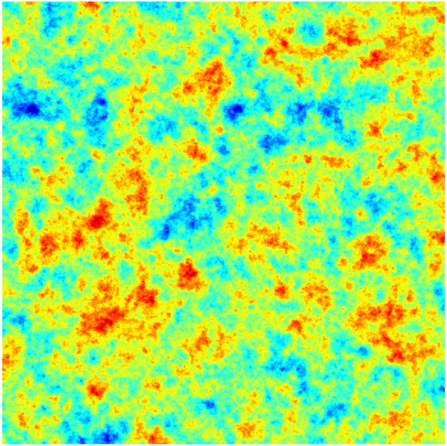
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

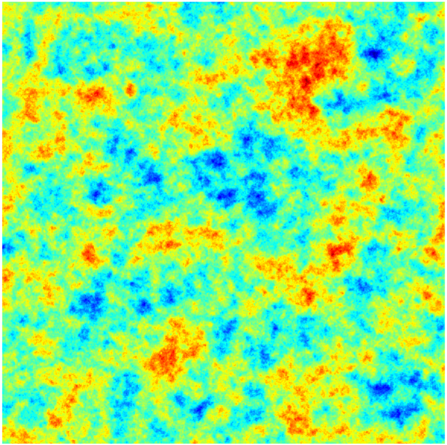
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

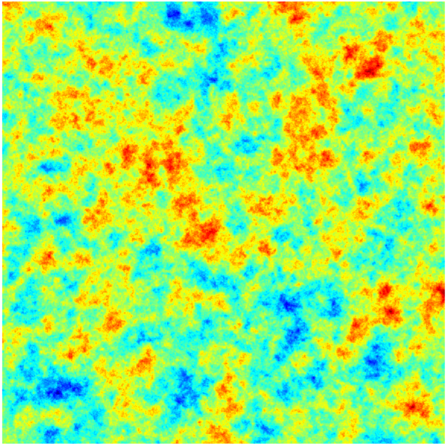
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

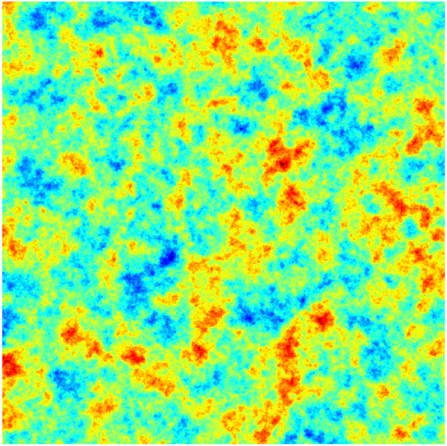
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

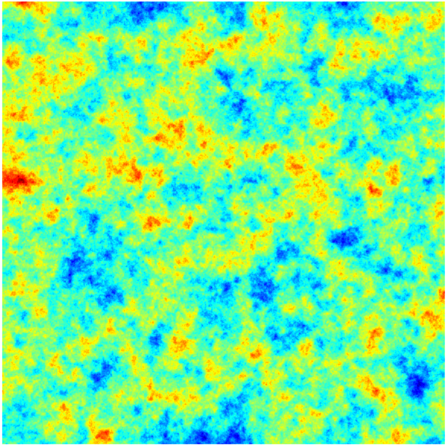
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

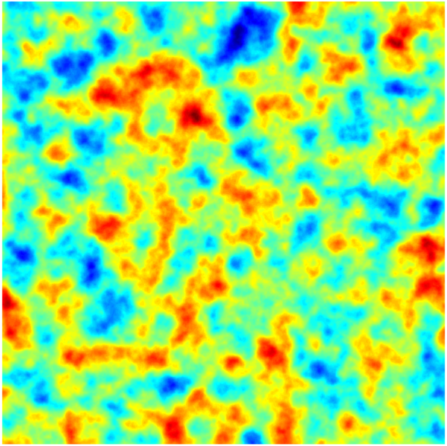
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

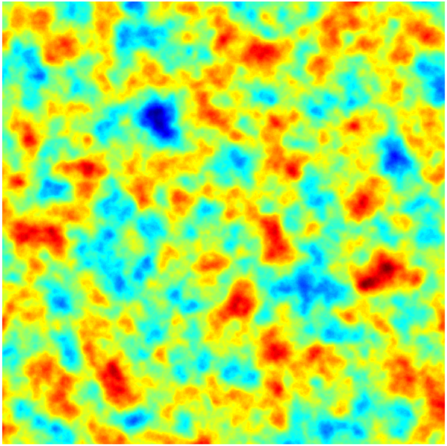
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

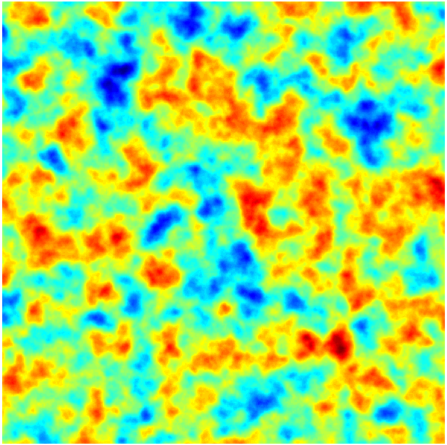
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

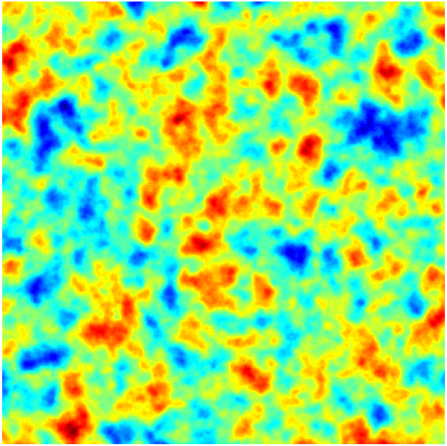
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

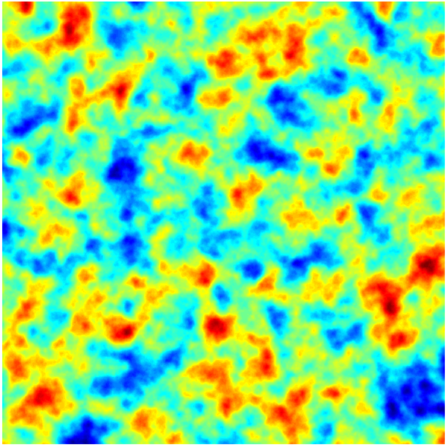
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

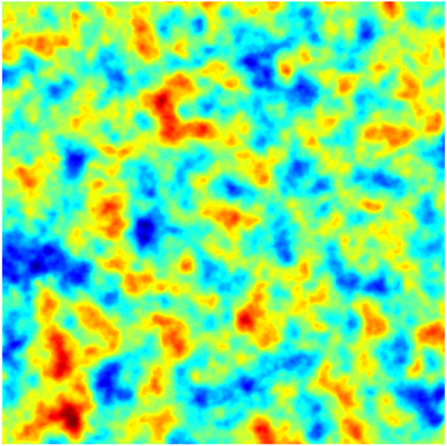
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

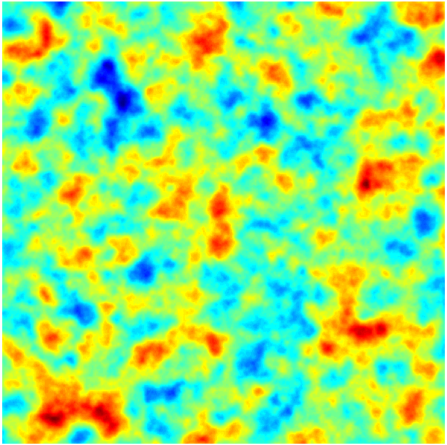
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

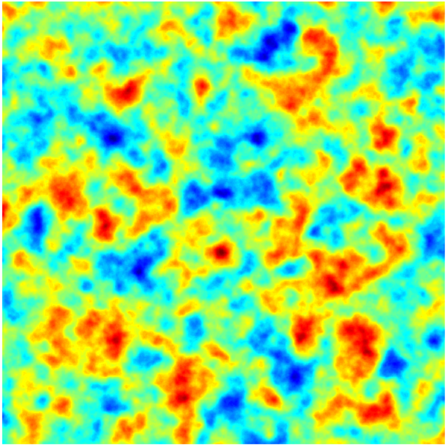
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

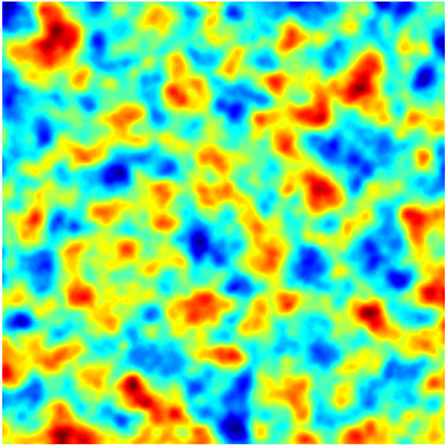
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

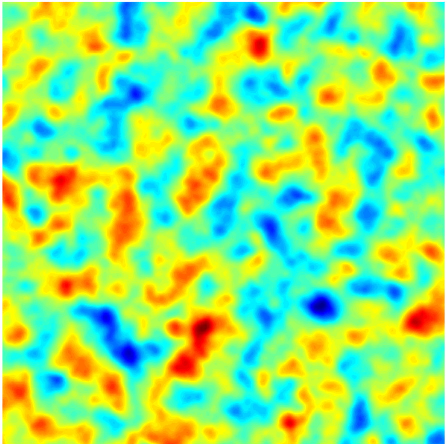
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

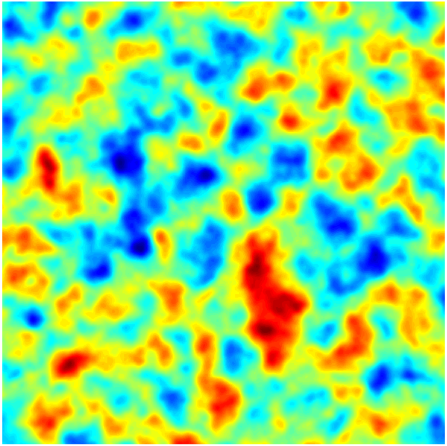
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

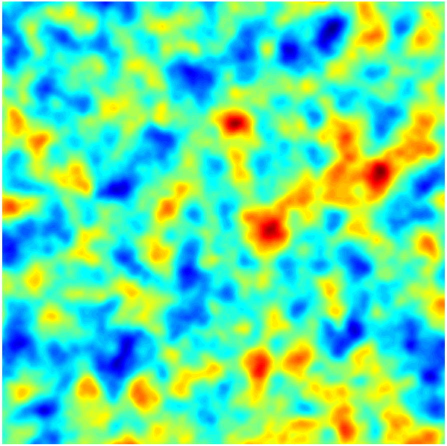
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

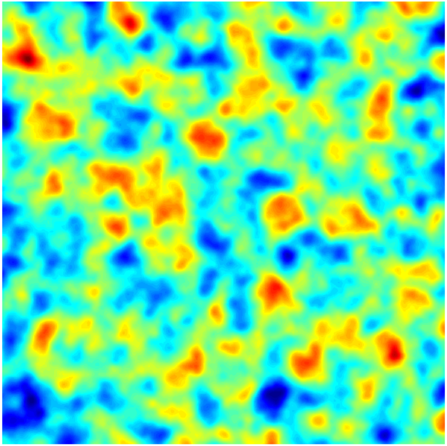
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

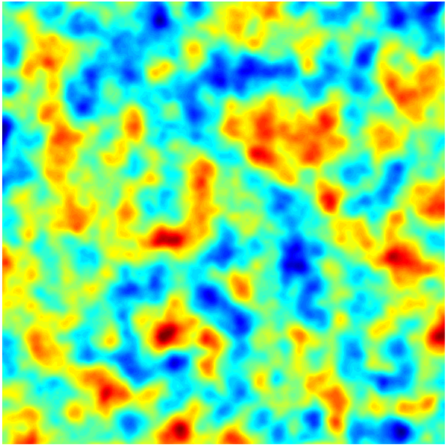
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

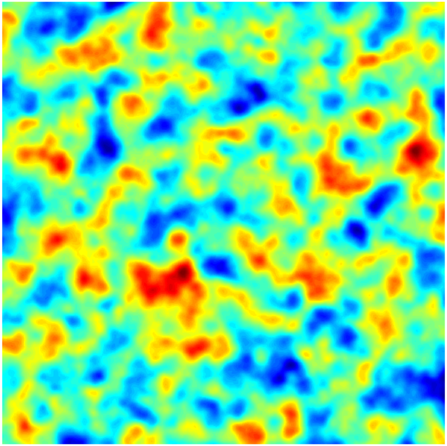
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

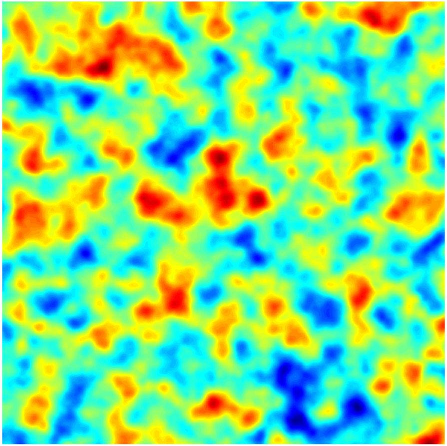
Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\ell = 0.05$

Covariance Eigenvalue Decay

Realizations of Gaussian RF



Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\ell = 0.05$

Contents

1. Introduction

2. Monte Carlo Methods

3. Random Fields

4. Stochastic Collocation

Contents

Stochastic Collocation

Introduction

Collocation methods are a long-established technique for solving integral or differential equations and are based on requiring the equation under consideration to hold at a finite number of collocation points sufficient to determine an approximate solution in an appropriate finite-dimensional function space.

They were introduced for solving PDEs with random inputs in [Xiu & Hesthaven, 2005] and [Babuška, Nobile & Tempone, 2007] and offer anumber of atractive features:

- Like MC, they reduce to a series of uncoupled deterministic subproblems for which legacy code can be used essentialy unmodified.
- Unlike MC, collocation can take advantage of smooth dependence of the solution on the random parameters to yield spectral convergence.
- Nonlinear problems pose no additional difficulty.

Stochastic Collocation

Setting

We consider the model problem on the bounded domain $D \subset \mathbb{R}^d$

$$-\nabla\cdot(a\nabla u) = f \quad \text{on } D, \qquad u|_{\partial D} = 0 \tag{4.1}$$

with random field data $\{a(\boldsymbol{x}), \boldsymbol{x} \in D\}$ and (possibly) $\{f(\boldsymbol{x}), \boldsymbol{x} \in D\}$.

We make the following assumptions:

Assumption 4.1

(a) $f \in L^2(\Omega; L^2(D))$.

(b) a is uniformly bounded from below, i.e., there exists a constant $a_{\min} > 0$ such that

$$a(\boldsymbol{x}) \geqslant a_{\min} \qquad \forall \boldsymbol{x} \in \overline{D}, \quad \mathbf{P}\text{-a.s.}$$

Stochastic Collocation

Setting

In addition to the space $\mathscr{V} := L^2(\Omega; H_0^1(D)) = L^2(\Omega;\mathbb{R}) \otimes H_0^1(D)$, we introduce the stochastic energy space

$$\mathscr{V}_a := \left\{ v \in \mathscr{V} : \|v\|_a := \mathbf{E}\left[(a\nabla v, \nabla v)_{L^2(D)} \right]^{1/2} < \infty \right\}.$$

Proposition 4.2

Under these assumptions $\mathscr{V}_a$ is continuously embedded in $\mathscr{V}$ and

$$\|v\|_{L^2(\Omega;H_0^1(D))} \leqslant \frac{1}{a_{\min}} \|v\|_{\mathscr{V}_a}.$$

Stochastic Collocation

Stochastic variational problem

With these definitions we give the following variational formulation of problem (4.1)

Find $u \in \mathscr{V}_a$ such that

$$\mathbf{E}\left[(a\nabla u, \nabla v)_{L^2(D)}\right] = \mathbf{E}\left[(f, v)_{L^2(D)}\right] \quad \forall v \in \mathscr{V}_a. \tag{4.2}$$

Lemma 4.3

Under Assumption 4.1, the variational problem (4.2) possesses a unique solution $u \in \mathscr{V}_a$ such that

$$\|u\|_{L^2(\Omega;H_0^1(D))} \leqslant \frac{C_D}{a_{\min}} \|f\|_{L^2(\Omega;L^2(D))},$$

where C_D denotes the Poincaré-Friedrichs constant of D.

Stochastic Collocation

Weaker assumptions on coefficient

If we assume the lower bound on the coefficient field a to hold only realization-wise, i.e.,

$$a(\boldsymbol{x}, \omega) \geqslant a_{\min}(\omega) > 0 \qquad \text{a.s. and a.e. on } D,$$

for a random variable $a_{\min}$, then Lemma 4.3 yields, for each $\omega \in \Omega$, a solution $u(\omega) \in H_0^1(D)$.

Lemma 4.4

Let $p, q \geqslant 0$ be conjugate exponents, i.e., $1/p + 1/q = 1$ and $k \in \mathbb{N}$. Then if $f \in L^{kp}(\Omega; L^2(D))$ and $1/a_{\min} \in L^{kq}(\Omega; \mathbb{R})$, we have $u \in L^k(\Omega; H_0^1(D))$.

Example 4.5

Lognormal Gaussian field

$$a(\boldsymbol{x}, \omega) = \exp\left(\sum_{m=1}^{M} g_m(\boldsymbol{x}) \xi_m(\omega)\right), \qquad \xi_m \text{ i.i.d., } \xi_m \sim \mathrm{N}(0, 1).$$

Stochastic Collocation

Finite-dimensional noise

Assumption 4.6 (Finite-dimensional noise)

The coefficient and source term in (4.1) have the form

$$a(\boldsymbol{x},\omega) = a(\boldsymbol{x},\xi_1(\omega),\dots,\xi_M(\omega)), \quad f(\boldsymbol{x},\omega) = f(\boldsymbol{x},\boldsymbol{x},\xi_1(\omega),\dots,\xi_M(\omega))$$

with $M \in \mathbb{N}$ and real-valued random variables $\{\xi_m\}$ with mean zero and unit variance. We denote by $\Gamma_m = \xi_m(\Omega)$ the image of each ξ_m, $\Gamma := \prod_{m=1}^{M} \Gamma_m$ and assume that the random vector $\boldsymbol{\xi} = [\xi_1,\dots,\xi_M]$ has a joint pdf

$$\rho : \Gamma \to \mathbb{R}_0^+ \qquad \text{with} \qquad \rho \in L^\infty(\Gamma).$$

- An example of such a situation is a random field represented as a truncated KL expansion.
- Typically f and a are assumed independent, i.e., the first depends on a random vector $\boldsymbol{\xi}_a(\omega)$ and the second on $\boldsymbol{\xi}_f(\omega)$ with both random vectors independent.

Stochastic Collocation

Parametric problem

The stochastic variational problem (4.2) may now be reformulated as a (deterministic) parametrized PDE with respect to the space

$$\mathscr{V}_{\rho,a} := L^2(\Gamma, \mathfrak{B}(\Gamma), \rho\,\mathrm{d}\boldsymbol{\xi}; H_0^1(D))$$

in place of $\mathscr{V}_a$:

Find $u \in \mathscr{V}_{a,\rho}$ such that

$$\int_\Gamma (a\nabla u, \nabla v)_{L^2(D)}\, \rho(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} = \int_\Gamma (f, v)_{L^2(D)}\, \rho(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} \qquad \forall v \in \mathscr{V}_{\rho,a}. \tag{4.3}$$

The solution then also has the form $u = u(\boldsymbol{x}, \boldsymbol{\xi}) \in \mathscr{V}_{\rho,a}$ with $\boldsymbol{x} \in D$, $\boldsymbol{\xi} \in \Gamma$. It is convenient to view u as a mapping

$$u : \Gamma \to H_0^1(D).$$

Contents

Stochastic Collocation

Basic idea

To approximate a parameter-dependent object $u = u(\boldsymbol{\xi})$ with values in an abstract space V, fix a finite-dimensional subspace $V_N = \operatorname{span}\{u_1, \dots, u_N\} \subset V$ and set

$$u(\boldsymbol{\xi}) \approx u_N(\boldsymbol{\xi}) = \sum_{j=1}^{N} u_j \, \psi_j(\boldsymbol{\xi})$$

with coefficient functions $\psi_j : \Gamma \to \mathbb{R}$ determined by a fixed set of

collocation points $\quad \{\boldsymbol{\xi}_j\}_{j=1}^N \subset \Gamma.$

Simplest choice for ψ_j: Lagrange basis of multivariate (global) polynomials with respect to a system

$$\boldsymbol{\Xi} := \{\boldsymbol{\xi}_j\}_{j=1}^N \subset \Gamma$$

of unisolvent nodes.

Stochastic Collocation

Lagrange interpolant

Given a univariate nodal sequence of distinct nodes

$$\chi_k = \{\xi_1^{(k)}, \dots, \xi_{n_k}^{(k)}\}, \qquad k \in \mathbb{N},$$

we denote by $\{\ell_j^{(k)}\}_{j=1}^{n_k}$ with $\ell_j^{(k)} \in \mathscr{P}_{n_k-1}$ the associated Lagrange basis, i.e., the uniquely determined polynomials of degree $n_k - 1$ satisfying

$$\ell_j^{(k)}(\xi_i^{(k)}) = \delta_{i,j}, \qquad j = 1, \dots, n_k.$$

We introduce the univariate interpolation operator

$$I_k : f \mapsto I_k f = \sum_{j=1}^{n_k} f(\xi_j^{(k)})\, \ell_j^{(k)} \in \mathscr{P}_{n_k-1}$$

Stochastic Collocation

Interpolation nodes

- We will later analyze tensor-product interpolation in the variable ξ and its approximation properties, which can be derived from the constituent univariate interpolations.
- For univariate interpolation, good nodal sequences are, e.g., zeros of orthogonal polynomials, Clenshaw-Curtiss nodes (extremal values of the Chebyshev polynomials) and Leja points.
- We will restrict ourselves to zeros of orthogonal polynomials. Since these are, at the same time, the nodes of high-order quadrature schemes, this will simplify the computation of integrals involving the collocation approximation, e.g., to compute moments of the solution of (4.1).

- If we assume Γ is the M-fold Cartesian product of the same (bounded or unbounded) real interval. In this case we may choose the same nodal sequence in all coordinates, and set

  $$\boldsymbol{\Xi}_k := \chi_k \times \cdots \times \chi_k = \{\boldsymbol{\xi}_{\boldsymbol{\alpha}} = (\xi^{(k)}_{\alpha_1}, \ldots, \xi^{(k)}_{\alpha_M}) : 1 \leqslant \alpha_m \leqslant n_k\}.$$

  Note that $N := |\boldsymbol{\Xi}_k| = n_k^M$.
- The tensor-product interpolation operator is then defined as

  $$\mathscr{I}_k := I_k \otimes \cdots \otimes I_k : u \mapsto \sum_{|\boldsymbol{\alpha}|_\infty \leqslant n_k} u(\boldsymbol{\xi}_{\boldsymbol{\alpha}})\, \ell^{(k)}_{\alpha_1} \cdot \ldots \cdot \ell^{(k)}_{\alpha_M},$$

  where $|\boldsymbol{\alpha}|_\infty = \max_{m=1}^M |\alpha_m|$.
- The range of the tensor-product interpolation operator $\mathscr{I}_k$ is the space $\mathscr{Q}_{n_k-1,M}$ of multivariate polynomials of degree $n_k - 1$ defined as

  $$\mathscr{Q}_{p,M} = \left\{ \prod_{m=1}^{M} p_m(\xi_m) : p_m \in \mathscr{P}_p \right\}.$$

Stochastic Collocation

Semi-discrete problem

- The semi-discrete problem is obtained by replacing $V = H_0^1(D)$ with a finite-dimensional subspace, say, a finite-element space $V^h \subset H_0^1(D)$.
- If we require the discrete variational problem to hold pointwise in Γ, we obtain the problem

Find $u^h : \Gamma \to V^h$ such that

$$(a(\boldsymbol{\xi})\nabla u(\boldsymbol{\xi}), \nabla v)_{L^2(D)} = (f(\boldsymbol{\xi}), v)_{L^2(D)} \qquad \forall v \in V^h \text{ and } \forall \boldsymbol{\xi} \in \Gamma. \tag{4.4}$$

Stochastic Collocation

Fully discrete problem

The fully discrete problem is obtained by approximating the semidiscrete solution $u^h : \Gamma \to V^h$ by

$$u^h(\boldsymbol{x},\boldsymbol{\xi}) \approx u^{h,p}(\boldsymbol{x},\boldsymbol{\xi}) := (\mathscr{I}_p u^h)(\boldsymbol{x},\boldsymbol{\xi}),$$

where $\mathscr{I}_p$ is the tensor-product interpolant constructed from univariate Lagrange interpolants of degree p, i.e., based on $p+1$ disctinct nodes in each variable.

This entails solving a (deterministic) version of (4.1) for each of the tensor-product interpolation nodes:

Find $u(\boldsymbol{\xi}_{\boldsymbol{\alpha}}) \in V^h$ for all $\boldsymbol{\xi}_{\boldsymbol{\alpha}} \in \boldsymbol{\Xi}$ such that

$$(a(\boldsymbol{\xi}_{\boldsymbol{\alpha}})\nabla u(\boldsymbol{\xi}_{\boldsymbol{\alpha}}), \nabla v)_{L^2(D)} = (f(\boldsymbol{\xi}_{\boldsymbol{\alpha}}), v)_{L^2(D)} \qquad \forall v \in V^h. \tag{4.5}$$

Stochastic Collocation

Auxiliary density

- We have not made the assumption that the random variables $\{\xi_m\}_{m=1}^M$ are independent. Expansions containing non-independent random variables arise naturally when other expansion functions than the covariance eigenfunctions are employed.
- Both analysis and computation, however, are considerably simplified when independence holds. To this end we introduce an auxiliary density function $\hat{\rho} : \Gamma \to \mathbb{R}_0^+$ with the properties

$$\hat{\rho}(\boldsymbol{\xi}) = \prod_{m=1}^{M} \hat{\rho}_m(\xi_m) \ \forall \boldsymbol{\xi} \in \Gamma \quad \text{and} \quad \left\| \frac{\rho}{\hat{\rho}} \right\|_{L^\infty(\Gamma)} < \infty. \tag{4.6}$$

- Since the density separates, it can be viewed as the joint pdf of M independent random variables.
- We choose as interpolation nodes the tensor product of univariate nodal sets consisting of the zeros of the orthogonal polynomials associated with the weight function $\hat{\rho}_m(\xi_m)$ in each of the M coordinates $\xi_1, \dots, \xi_M$.

Contents

1. Introduction

2. Monte Carlo Methods

3. Random Fields

4. Stochastic Collocation

4.1 Introduction
4.2 Collocation
4.3 Analytic Parameter Dependence
4.4 Convergence

Stochastic Collocation

Weighted space

Our analysis requires assumptions on f and the densities $\hat{\rho}$ and ρ:

- f is a continuous function of $\boldsymbol{\xi}$ which, in case of unbounded parameter domain Γ, grows at most exponentially at infinity.
- ρ and $\hat{\rho}$ behave at infinity like a Gaussian density.

To make these assumptions explicit we introduce a weight function

$$\sigma(\boldsymbol{\xi}) := \prod_{m=1}^{M} \sigma_m(\xi_m) \leqslant 1, \quad \sigma_m(\xi_m) = \begin{cases} 1 & \text{if } \Gamma_m \text{ bounded,} \\ e^{-\alpha_m |\xi_m|} & \text{otherwise,} \end{cases} \tag{4.7}$$

as well as the function space

$$C_\sigma(\Gamma; W) := \left\{ v : \Gamma \to W,\ v \text{ continuous in } \boldsymbol{\xi},\ \max_{\boldsymbol{\xi} \in \Gamma} \|\sigma(\boldsymbol{\xi}) v(\boldsymbol{\xi})\|_W < \infty \right\}$$

where W is a Banach space of functions defined on D.

Assumption 4.7 (Growth at infinity)

In what follows we assume that

(a) $f \in C_\sigma(\Gamma; L^2(D))$ and

(b) the joint probability density ρ satisfies

$$\rho(\boldsymbol{\xi}) \leqslant C_\rho \, e^{-\sum_{m=1}^{M} (\delta_m \xi_m)^2} \quad \forall \boldsymbol{\xi} \in \Gamma \tag{4.8}$$

for some $C_\rho > 0$ and $\delta_m > 0$ if Γ_m is unbounded and $\delta_m = 0$ otherwise.

- We can now choose any suitable auxiliary density $\hat{\rho}(\boldsymbol{\xi}) = \prod_{m=1}^{M} \hat{\rho}_m(\xi_m)$ that satisfies, for each $m = 1, \dots, M$,

$$C_{\min}^{(m)} e^{-(\delta_m \xi_m)^2} \leqslant \hat{\rho}_m(\xi_m) \leqslant C_{\max}^{(m)} e^{-(\delta_m \xi_m)^2}, \quad \forall \xi_m \in \Gamma_m, \tag{4.9}$$

  for positive constants $C_{\min}^{(m)}$, $C_{\max}^{(m)}$ independent of ξ_m.
- This choice satisfies the requirement (4.6) with

$$\left\| \frac{\rho}{\hat{\rho}} \right\|_{L^\infty(\Gamma)} \leqslant \frac{C_\rho}{C_{\min}}, \qquad C_{\min} := \prod_{m=1}^{M} C_{\min}^{(m)}.$$

- Under the above assumptions we have the inclusions

$$C_\sigma(\Gamma; W) \subset L^2_{\hat{\rho}}(\Gamma; W) \subset L^2_\rho(\Gamma; W)$$

  with continuous imbeddings. $(L^2_\rho(\Gamma; W) := L^2(\Gamma, \mathfrak{B}(\Gamma), \rho(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}; W))$

Stochastic Collocation

Weighted space

Lemma 4.8

If $f \in C_\sigma(\Gamma; L^2(D))$ and $a \in C_{\mathsf{loc}}(\Gamma; L^\infty(D))$, uniformly bounded away from zero, then the solution to problem (4.3) satisfies $u \in C_\sigma(\Gamma; H_0^1(D))$.

Stochastic Collocation

Analytic extension

We next show that, if a and f possess partial derivatives of all orders with respect to $\boldsymbol{\xi}$ with mild growth, then the solution u is analytic as a function of each individual parameter ξ_m. This requires a one-dimensional analysis, for which we introduce the following notation:

$$\Gamma_m^* := \bigtimes_{\substack{j=1,\\ j\neq m}}^{M} \Gamma_j \quad \text{with generic elements denoted } \boldsymbol{\xi}_m^*, \quad m = 1, \dots, M.$$

Similarly, we set

$$\hat{\rho}_m^* := \prod_{\substack{j=1,\\ j\neq m}}^{M} \hat{\rho}_j \quad \text{and} \quad \sigma_m^* := \prod_{\substack{j=1,\\ j\neq m}}^{M} \sigma_j \ .$$

Stochastic Collocation

Analytic extension

Lemma 4.9

Under the assumption that, for every $\boldsymbol{\xi} = (\xi_m, \boldsymbol{\xi}_m^*) \in \Gamma$, there exists $\gamma_m < \infty$ such that

$$\left\| \frac{\partial_{\xi_m}^k a(\boldsymbol{\xi})}{a(\boldsymbol{\xi})} \right\|_{L^\infty(D)} \leqslant \gamma_m^k \, k! \qquad \text{and} \qquad \frac{\|\partial_{\xi_m}^k f(\boldsymbol{\xi})\|_{L^2(D)}}{1 + \|f(\boldsymbol{\xi})\|_{L^2(D)}} \leqslant \gamma_m^k \, k!, \tag{4.10}$$

the solution $u(\xi_m, \boldsymbol{\xi}_m^*)$ as a function of ξ_m. $u : \Gamma_m \to C_{\sigma_m^*}(\Gamma_m^*; H_0^1(D))$ admits an analytic continuation $u(\zeta, \boldsymbol{\xi}_m^*)$, $\zeta \in \mathbb{C}$, to the region of the complex plane

$$\Sigma(\Gamma_m; \tau_m) := \{\zeta \in \mathbb{C} : \operatorname{dist}(\zeta, \Gamma_m) \leqslant \tau_m\} \tag{4.11}$$

with $0 < \tau_m < 1/(2\gamma_m)$. Moreover, for all $\zeta \in \Sigma(\Gamma_m; \tau_n)$ there holds

$$\|\sigma_m(\operatorname{Re}\zeta) u(\zeta)\|_{C^*_{\sigma_m}(\Gamma_m^*; H_0^1(D))} \leqslant \frac{C_D}{a_{\min}} \frac{e^{\alpha_m \tau_m}}{1 - 2\tau_m \gamma_m} (1 + 2\|f\|_{C_\sigma(\Gamma; H_0^1(D))}). \tag{4.12}$$

Stochastic Collocation

Analytic extension, examples

If the diffusion coefficient is expanded in a finite linear KL series

$$a(\boldsymbol{x},\omega) = \overline{a}(\boldsymbol{x}) + \sum_{m=1}^{M} \sqrt{\lambda_m}\, a_m(\boldsymbol{x})\, \xi_m(\omega),$$

assuming $a(\boldsymbol{x},\omega) \geqslant a_{\min}$ **P**-a.s. and a.e. in D, then we have the bounds

$$\left\| \frac{\partial_{\xi_m}^k a}{a} \right\|_{L^\infty(\Gamma\times D)} \leqslant \begin{cases} \dfrac{\sqrt{\lambda_m}\|a_m\|_{L^\infty(D)}}{a_{\min}}, & k = 1, \\ 0, & k > 1. \end{cases}$$

and we may choose

$$\gamma_m = \frac{\sqrt{\lambda_m}\|a_m\|_{L^\infty(D)}}{a_{\min}}$$

in (4.10).

Stochastic Collocation

Analytic extension, examples

If the diffusion coefficient is expanded in a finite exponential KL series

$$a(\boldsymbol{x},\omega) = a_{\min} + \exp\left(\overline{a}(\boldsymbol{x}) + \sum_{m=1}^{M} \sqrt{\lambda_m}\, a_m(\boldsymbol{x})\, \xi_m(\omega)\right)$$

we have

$$\left\| \frac{\partial_{\xi_m}^k a}{a} \right\|_{L^\infty(\Gamma\times D)} \leqslant \left(\sqrt{\lambda_m} \|a_m\|_{L^\infty(D)}\right)^k$$

and we can set

$$\gamma_m = \sqrt{\lambda_m} \|a_m\|_{L^\infty(D)}$$

in (4.10).

Stochastic Collocation

Analytic extension, examples

If the source term f has the form

$$f(\boldsymbol{x},\omega) = \overline{f}(\boldsymbol{x}) + \sum_{m=1}^{M} f_m(\boldsymbol{x})\,\xi_m(\omega)$$

with Gaussian RV ξ_m (not necessarily independent) and the functions f_m are square integrable, then f belongs to $C_\sigma(\Gamma; L^2(D))$ with weight σ as defined in (4.7) for any choice of exponential coefficients $\alpha_m > 0$.

Moreover

$$\frac{\|\partial^k_{\xi_m} f(\boldsymbol{\xi})\|_{L^2(D)}}{1 + \|f(\boldsymbol{\xi})\|_{L^2(D)}} \leqslant \begin{cases} \|f_m\|_{L^2(D)}, & k = 1, \\ 0, & k > 1, \end{cases}$$

and we can take $\gamma = \|f_m\|_{L^2(D)}$ in (4.10).

Thus such a source term satisfies the assumptions of Lemma 4.9.

Note also that, in case a is deterministic, the solution u is linear in the ξ_m, and hence clearly analytic.

Contents

Stochastic Collocation

Convergence

We collect some classical results on interpolation theory and consider univariate functions f defined on a bounded or unbounded interval $\Gamma \subset \mathbb{R}$ with values in a Hilbert space V.

- As before, assume ρ is a positive weight function on Γ which satisfies

  $$\rho(\xi) \leqslant C_M\, e^{-(\delta\xi)^2} \qquad \text{for some } C_M > 0$$

  and $\delta > 0$ for unbounded Γ and $\delta = 0$ otherwise.
- We let $\{\vartheta_j\}_{j=1}^{p+1}$ denote the zeros of the orthogonal polynomial of degree $p+1$ associated with the weight function ρ.
- Let σ be an additional positive weight function such that

  $$\sigma(\xi) \geqslant C_m\, e^{-(\delta\xi)^2/4} \qquad \text{for some } C_m > 0.$$
- Observe that the condition on σ is satisfied both by a Gaussian weight $\sigma(\xi) = e^{-(\mu\xi)^2}$ with $\mu \leqslant \delta/2$ and by an exponential weight $\sigma(\xi) = e^{-\alpha|\xi|}$ for any $\alpha \geqslant 0$.

Lemma 4.10

Let $\Gamma \subset \mathbb{R}$ be an interval (bounded or unbounded) and let $\rho : \Gamma \to \mathbb{R}^+$ denote a weight function such that all integer moments are finite, i.e., $\int_\Gamma \xi^n \, \rho(\xi) \, \mathrm{d}\xi < \infty$, $n \in \mathbb{N}_0$. Then for each $p \in \mathbb{N}$ there exist polynomials $\{q_j\}_{j=1}^{p+1}$ of degree p such that for all $1 \leqslant j, k \leqslant p+1$ there holds

$$\begin{aligned} (q_j, q_k)_\rho &:= \int_\Gamma q_m(\xi) q_n(\xi) \, \rho(\xi) \, \mathrm{d}\xi = \delta_{j,k}, \\ \text{and} \qquad (q_j, q_k)_{\tilde{\rho}} &= \vartheta_j \delta_{j,k,n}, \end{aligned} \tag{4.13}$$

where $\tilde{\rho}(\xi) := \xi \rho(\xi)$. Moreover, the q_j are, up to a constant factor, the Lagrange basis polynomials $\{\ell_j\}_{j=1}^{p+1}$ constructed with the $p+1$ (distinct) zeros of the orthogonal polynomial of degree $p+1$ associated with the weight function ρ.

The ϑ_j are the nodes of the associated $(p+1)$-point Gauss quadrature rule with weights given by

$$\omega_j = \int_\Gamma \ell_j(\xi) \rho(\xi) \, \mathrm{d}\xi = \int_\Gamma \ell_j(\xi)^2 \rho(\xi) \, \mathrm{d}\xi, \quad j = 1, \ldots, p+1.$$

Stochastic Collocation

Convergence

By $I_p : C(\Gamma) \to \mathscr{P}_p$ we denote the Lagrange interpolation operator

$$(I_p f)(\xi) = \sum_{j=1}^{p+1} f(\vartheta_j)\ell_j(\xi), \qquad \xi \in \Gamma.$$

Lemma 4.11

The operator $I_p : C_\sigma(\Gamma, V) \to L^2_\rho(\Gamma; V)$ is continuous.

Lemma 4.12

For every function $v \in L^2_\rho(\Gamma; V)$ the interpolation error satisfies

$$\|v - I_p v\|_{L^2_\rho(\Gamma;V)} \leqslant C \inf_{w \in \mathscr{P}_p \otimes V} \|v - w\|_{C_\sigma(\Gamma;V)}$$

with a constant C independent of p.

Stochastic Collocation

Convergence

Lemma 4.13

Given a function $v \in C(\Gamma; V)$ which admits an analytic extension to the region

$$\Sigma(\Gamma;\tau) := \{z \in \mathbb{C} : \operatorname{dist}(z,\Gamma) \leqslant \tau\}$$

of the complex plane for some $\tau > 0$, then there holds

$$\min_{w \in \mathscr{P}_p \otimes V} \|v - w\|_{C(\Gamma;V)} \leqslant \frac{2}{\rho - 1} e^{-p \log \rho} \max_{z \in \Sigma(\Gamma;\tau)} \|v(z)\|_V,$$

where

$$\rho := \frac{2\tau}{|\Gamma|} + \sqrt{1 + \frac{4\tau^2}{|\Gamma|^2}} \geqslant 1.$$

A proof can be found in [Babuška et al., 2007] Lemma 4.4 and general results on best approximation of analytic functions by polynomials in [DeVore & Lorentz, 1993] Chapter 7, Section 8.

Stochastic Collocation

Convergence

In case of unbounded Γ we recall a theorem of [Hille, 1940] on the convergence of Hermite series and the decay of the associated expansion coefficients.

- Let $H_n \in \mathscr{P}_n$ denote the (univariate) Hermite polynomial of degree n

$$H_n(\xi) = \frac{(-1)^n}{\sqrt{\pi^{1/2} 2^n n!}} \, e^{\xi^2} \frac{\mathrm{d}^n}{\mathrm{d}\xi^n} e^{-\xi^2}, \qquad n \in \mathbb{N}_0,$$

  and by $h_n(\xi) = e^{-\xi^2/2} H_n(\xi)$ the associated Hermite function.
- The Hermite polynomials are orthogonal on $\mathbb{R}$ with respect to the weight function $e^{-\xi^2}$ and form a complete orthonormal system of $L^2(\mathbb{R})$ with respect to the associated inner product.
- The Hermite polynomials and functions as defined above are normalized in such a way that

$$\int_{\mathbb{R}} h_k(\xi) h_\ell(\xi) \, \mathrm{d}\xi = \int_{\mathbb{R}} H_k(\xi) H_\ell(\xi) e^{-\xi^2} \, \mathrm{d}\xi = \delta_{k,\ell}, \qquad k, \ell \in \mathbb{N}_0.$$

Stochastic Collocation

Convergence

Lemma 4.14 (Hille, 1940)

Let the function f be analytic in the strip $\{|\operatorname{Im} z| \leqslant \tau\}$. A necessary and sufficient condition for the Fourier-Hermite series

$$\sum_{k=0}^{\infty} f_k \, h_k(z), \qquad f_k := \int_{\mathbb{R}} f(\xi)\, h_k(\xi)\, \mathrm{d}\xi, \tag{4.14}$$

to converge to $f(z)$ in $\Sigma(\mathbb{R}; \tau)$ is that for every $\beta \in [0, \tau)$ there exist a finite positive $C(\beta)$ such that

$$|f(x+iy)| \leqslant C(\beta) e^{-|x|\sqrt{\beta^2 - y^2}}, \qquad x \in \mathbb{R}, |y| \leqslant \beta. \tag{4.15}$$

Moreover, the Fourier coefficients satisfy

$$|f_k| \leqslant C e^{-\tau\sqrt{2k+1}}. \tag{4.16}$$

Stochastic Collocation

Convergence

Lemma 4.15

Assume that $v \in C_\sigma(\mathbb{R}; V)$ admits an analytic extension to the strip

$$\Sigma(\mathbb{R};\tau) = \{z \in \mathbb{C} : \operatorname{dist}(z,\mathbb{R}) \leqslant \tau\} \qquad \text{for some} \qquad \tau > 0$$

and that

$$\sigma(x)\|v(z)\|_V \leqslant C_v(\tau) \qquad \forall z = x + iy \in \Sigma(\mathbb{R};\tau).$$

Then for any $\delta > 0$ there exists a constant C independent of p and a function $\Theta(p) = O(p)$ such that

$$\min_{w \in \mathscr{P}_p \otimes V} \max_{\xi \in \mathbb{R}} \left| \|v(\xi) - w(\xi)\|_V \, e^{-(\delta\xi)^2/4} \right| \leqslant C \, \Theta(p) \, e^{-\tau\delta\sqrt{p}}.$$

Theorem 4.16

Under the assumptions of Lemmas 4.8 and 4.9 there exist positive constants $\{r_m\}_{m=1}^M$ and C independent of h and p such that

$$\|u - u^{h,p}\|_{L^2_\rho(\Gamma,V)} \leqslant \frac{1}{\sqrt{a_{\min}}} \inf_{v \in L^2_\rho(\Gamma,V^h)} \|u - v\|_{\mathscr{V}_a} + C \sum_{m=1}^M \beta_m(p_m) \exp(-r_m p_m^{\theta_m}) \qquad (4.17)$$

where, if Γ_m is bounded,

$$\theta_m = \beta_m = 1, \qquad r_m = \log\left[\frac{2\tau_m}{|\Gamma_m|}\left(1 + \sqrt{1 + \frac{|\Gamma_m|^2}{4\tau_m^2}}\right)\right]$$

and, if Γ_m is unbounded,

$$\theta_m = 1/2, \qquad \beta_m = O(\sqrt{p_m}), \qquad r_m = \tau_m \delta_m.$$

τ_m is smaller than the distance between Γ_m and the nearest singularity of u as defined in Lemma 4.9 and δ_m is as defined in (4.8).

Contents

5 Probability Theory

6 Elliptic Boundary Value Problems

7 Collection of Results from Functional Analysis

8 Miscellanea

Contents

We denote an abstract probability space by $(\Omega, \mathfrak{A}, \mathbf{P})$, in which

Ω is an abstract set of elementary events,

$\mathfrak{A}$ is a σ-algebra of subsets of Ω containing the measurable events and

P is a probability measure on $\mathfrak{A}$.

Definition A.1

A measure **P** on a measurable space $(\Omega, \mathfrak{A})$ is called a probability measure if $\mathbf{P}(\Omega) = 1$.

Definition A.2

An event $A \in \mathfrak{A}$ is said to occur almost surely with respect to the measure **P** (**P**-a.s.) if $\mathbf{P}(A) = 1$.

Proposition A.3 (Boole's inequality)

For events $\{A_n\}_{n\in\mathbb{N}}$ there holds

$$\mathbf{P}\left(\cup_{n=1}^{\infty} A_n\right) \leqslant \sum_{n=1}^{\infty} \mathbf{P}(A_n).$$

Definition A.4

The set of all $\omega \in \Omega$ such that $\omega \in A_n$ for infinitely many values of n is defined as

$$\{A_n, \text{ i.o. }\} := \limsup_{n\in\mathbb{N}} A_n := \cap_{k=1}^{\infty} \cup_{n=k}^{\infty} A_n$$

Theorem A.5 (Borel-Cantelli Lemma)

If $\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$, then $\mathbf{P}\{A_n, \text{i.o.}\} = 0$. For independent events $\{A_n\}_{n\in\mathbb{N}}$ such that $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$ there holds $\mathbf{P}\{A_n, \text{i.o.}\} = 1$.

Probability Theory

Random variables

Definition A.6

Let $(\Omega, \mathfrak{A}, \mathbf{P})$ be a probability space and $(E, \mathfrak{E})$ a measurable space. A measurable function $X : \Omega \to E$ is called an (E-valued) random variable. Individual values $X(\omega)$ for $\omega \in \Omega$ are called realisations of the random variable.

Remark: If E is a topological space then the σ-algebra generated by the open subsets of E is called the Borel σ-algebra $\mathfrak{B}(E)$.

Definition A.7

Let X be an E-valued random variable where $(E, \mathfrak{E})$ is a measurable space and $(\Omega, \mathfrak{A}, \mathbf{P})$ is the underlying probability space. The probability distribution $\mathbf{P}_X$ of X (also called the law of X) is the probability measure on $(E, \mathfrak{E})$ defined by $\mathbf{P}_X(A) := \mathbf{P}(X^{-1}(A))$ for pre-images $X^{-1}(A) := \{\omega \in \Omega : X(\omega) \in A)\}$ of sets $A \in \mathfrak{E}$.

Remark: This construction is sometimes called the push-forward measure defined by $(\Omega, \mathfrak{A}, \mathbf{P})$, $(E, \mathfrak{E})$ and X.

Theorem A.8 (Doob-Dynkin lemma)

Let $f : \Omega \to E$ and $g : \Omega \to F$ be two measurable functions from a measurable space $(\Omega, \mathfrak{A})$ to two measurable spaces $(E, \mathfrak{E})$ and $(F, \mathfrak{F})$ of which the first is a separable and complete metric space. Then f is g-measurable if and only if there exists some measurable mapping $h : F \to E$ with $f = h \circ g$.

See [Kallenberg, 1997], Lemma 1.13 for a proof.

Definition A.9

The expectation of a Banach space-valued random variable X is defined as the integral

$$\mathbf{E}\left[X\right] := \int_\Omega X(\omega)\,\mathrm{d}\mathbf{P}(\omega).$$

Definition A.10

The k-th moment ($k \in \mathbb{N}$) of a real-valued random variable X is $\mathbf{E}\left[X^k\right]$.
The first moment $\mu := \mathbf{E}\left[X\right]$ is also called the mean or mean value.
The central moments $\mathbf{E}\left[(X-\mu)^k\right]$ measure the deviation of X from its mean.
The second central moment

$$\mathbf{Var}\,X := \mathbf{E}\left[(X-\mu)^2\right] = \mathbf{E}\left[X^2\right] - \mu^2$$

of a random variable X is called its variance.

Remark: The quantity $\sigma := \sqrt{\mathbf{Var}\,X}$ is called the standard deviation of X.

Probability Theory

Computation of moments

Moments of a random variable are sometimes more easily computed by integrating over the image variable.

Consider a real-valued random variable X from $(\Omega, \mathfrak{A})$ to $(\Gamma, \mathfrak{B}(\Gamma))$ where $\Gamma \subset \mathbb{R}$. For $B \in \mathfrak{B}(\Gamma)$, set $A := X^{-1}(B)$. Then by the definition of the probability distribution $\mathbf{P}_X$

$$\int_\Omega \mathbb{1}_A(\omega)\,\mathrm{d}\mathbf{P}(\omega) = \mathbf{P}(A) = \mathbf{P}_X(B) = \int_\Gamma \mathbb{1}_B(x)\,\mathrm{d}\mathbf{P}_X(x).$$

For measurable functions $f : \Gamma \to \mathbb{R}$ we have

$$\int_\Omega f(X(\omega))\,\mathrm{d}\mathbf{P}(\omega) = \int_\Gamma f(x)\,\mathrm{d}\mathbf{P}_X(x)$$

and, in particular,

$$\mathbf{E}\,[X] = \int_\Omega X(\omega)\,\mathrm{d}\mathbf{P}(\omega) = \int_\Gamma x\,\mathrm{d}\mathbf{P}_X(x).$$

Definition A.11

Let $\mathbf{P}$ be a probability measure on $(\Gamma, \mathfrak{B}(\Gamma))$ for some $\Gamma \subset \mathbb{R}$. If there exists a function $p : \Gamma \to [0, \infty)$ such that $\mathbf{P}(B) = \int_B p(x)\,\mathrm{d}x$ for any $B \in \mathfrak{B}(\Gamma)$ we say that $\mathbf{P}$ has a density p with respect to Lebesgue measure and we call p its probability density function (pdf). If X is a Γ-valued random variable on $(\Omega, \mathfrak{A}, \mathbf{P})$, the pdf p_X of X (if it exists) is the pdf of the probability distribution $\mathbf{P}_X$.

For real-valued random variables X from $(\Omega, \mathfrak{A}, \mathbf{P})$ to $(\Gamma, \mathfrak{B}(\Gamma))$ we then have[3]

$$\mathbf{E}[X] = \int_\Omega X(\omega)\,\mathrm{d}\mathbf{P}(\omega) = \int_\Gamma x\,\mathrm{d}\mathbf{P}_X(x) = \int_\Gamma xp(x)\,\mathrm{d}x. \qquad \text{(A.1)}$$

Event probabilities are then easily calculated as

$$\mathbf{P}(X \in (a,b)) = \mathbf{P}\left(\{\omega \in \Omega : a < X(\omega) < b\}\right) = \mathbf{P}_X((a,b)) = \int_a^b p(x)\,\mathrm{d}x.$$

[3] (where we have omitted the subscript X)

Probability Theory

Uniform distribution

A random variable X is uniformly distributed on $D = [a, b] \subset \mathbb{R}$, $(a < b)$, denoted

$$X \sim U(a, b),$$

if its pdf is

$$p(x) = \frac{1}{b-a}, \qquad x \in [a, b].$$

Using (A.1), we easily obtain

$$\mathbf{E}[X] = \int_a^b \frac{x}{b-a}\,\mathrm{d}x = \frac{a+b}{2}, \qquad \mathbf{E}\left[X^2\right] = \int_a^b \frac{x^2}{b-a}\,\mathrm{d}x = \frac{b^3-a^3}{3(b-a)},$$

so that $\mathbf{Var}\,X = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2 = \frac{(b-a)^2}{12}$.

Probability Theory

Gaussian distribution

A random variable X is said to follow the Gaussian or normal distribution on $\Gamma = \mathbb{R}$ if its pdf is given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R},$$

with two real parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, denoted $X \sim N(\mu, \sigma^2)$.
As is easily verified,

$$\mathbf{E}[X] = \mu, \qquad \mathbf{Var}\, X = \sigma^2.$$

The probability that X is within α of its mean is given by

$$\mathbf{P}(|X-\mu| \leqslant \alpha) = \operatorname{erf}\left(\frac{\alpha}{\sqrt{2\sigma^2}}\right),$$

with the error function erf defined by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, \mathrm{d}t.$$

The cumulative distribution function (cdf) of the standard normal distribution $\mathrm{N}(0,1)$ is denoted by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \, \mathrm{d}t = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right).$$

Any (finite) linear combination of (jointly) random variables is normally distributed.

Probability Theory

Change of variables formula

Lemma A.12 (Change of variables)

Suppose $Y : \Omega \to \mathbb{R}$ is a real-valued random variable and $f : (a, b) \to \mathbb{R}$ is continuously differentiable with inverse function f^{-1}. If p_Y is the pdf of Y, the pdf of the random variable $X : \Omega \to (a, b)$ defined via $X = f^{-1}(Y)$ is

$$p_X(x) = p_Y(f(x))\,|f'(x)| \qquad \text{for } a < x < b.$$

Probability Theory

Lognormal distribution

If $Y \sim N(\mu, \sigma^2)$, then the random variable

$$X := \exp(Y)$$

is said to follow a lognormal distribution. With $f(x) = \log x$, Lemma A.12 yields the pdf of X as

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2 x^2}} \exp\left(-\frac{[\log(x) - \mu]^2}{2\sigma^2}\right).$$

Moreover, there holds

$$\mathbf{E}\left[X\right] = \exp\left(\mu + \frac{\sigma^2}{2}\right), \qquad \mathbf{Var}\, X = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}.$$

Definition A.13

The covariance between two real-valued random variables is defined as

$$\mathbf{Cov}(X,Y) = \mathbf{E}\left[(X-\mu_X)(Y-\mu_Y)\right],$$

where $\mu_X := \mathbf{E}[X]$ and $\mu_Y := \mathbf{E}[Y]$. In particular, $\mathbf{Cov}(X,X) = \mathbf{Var}\, X$.

Note: An equivalent expression is $\mathbf{Cov}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y]$.

Calculation of the covariance requires evaluating the integral

$$\mathbf{E}[XY] = \int_\Omega X(\omega)Y(\omega)\,\mathrm{d}\mathbf{P}(\omega) = \int_{X(\Omega)\times Y(\Omega)} xy\,\mathrm{d}\mathbf{P}_{X,Y}(x,y),$$

in which $\mathbf{P}_{X,Y}$ is the joint probability distribution of X and Y.
Sometimes it is useful to scale the covariance to lie in $[-1,1]$. The resulting quantity is known as the correlation coefficient

$$\rho(X,Y) := \frac{\mathbf{Cov}(X,Y)}{\sigma_X\sigma_Y}.$$

Definition A.14

The joint probability distribution of two random variables X and Y is the distribution of the bivariate random variable $\boldsymbol{X} = (X, Y)$, i.e., for all $B \in \mathfrak{B}(X(\Omega) \times Y(\Omega))$

$$\mathbf{P}_{X,Y}(B) = \mathbf{P}(\{\omega \in \Omega : \boldsymbol{X}(\omega) \in B\}).$$

If it exists, the density $p_{X,Y}$ of $\mathbf{P}_{X,Y}$ is known as the joint pdf and

$$\mathbf{P}_{X,Y} = \int_B p_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y.$$

Definition A.15

If $\mathbf{Cov}(X, Y) = 0$ the random variables X and Y are said to be uncorrelated. A family $\{X_\alpha\}_\alpha$ is said to be pairwise uncorrelated if X_α and X_β are uncorrelated for all $\alpha \neq \beta$.

Note: Uncorrelated random variables may still be strongly related. As an example,

$$X \sim N(0,1), \qquad \text{and} \qquad Y := \cos X$$

satisfy $\mu_X = 0$ and hence

$$\begin{aligned} \mathbf{Cov}(X,Y) &= \mathbf{E}\left[X \cos X\right] = \int_{\mathbb{R}} x \cos(x)\, \mathrm{d}\mathbf{P}_X(x) \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x \cos(x) \exp\left(\frac{-x^2}{2}\right) \mathrm{d}x = 0. \end{aligned}$$

A stronger notion is that of independent random variables.

Probability Theory

Sub σ-algebras, σ-algebras generated by random variables

Definition A.16

A σ-algebra $\mathfrak{B}$ is a sub σ-algebra of $\mathfrak{A}$ if $\mathfrak{B} \subset \mathfrak{A}$, i.e., if $A \in \mathfrak{B}$ implies $A \in \mathfrak{A}$.

Definition A.17

Let X be an E-valued random variable on $(\Omega, \mathfrak{A}, \mathbf{P})$ for a measurable space $(E, \mathfrak{E})$. The σ-algebra generated by X, denoted $\sigma(X)$, is defined as

$$\sigma(X) := \{X^{-1}(A) : A \in \mathfrak{E}\} \subset \mathfrak{A}.$$

Remark: $\sigma(X)$ is the smallest σ-algebra such that X is measurable. It may be considerably smaller than $\mathfrak{A}$.

Definition A.18

Two events $A, B \in \mathfrak{A}$ are independent if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$.
Two σ-algebras $\mathfrak{A}_1$ and $\mathfrak{A}_2$ are independent if all pairs of events A_1 and A_2 with $A_1 \in \mathfrak{A}_1$ and $A_2 \in \mathfrak{A}_2$ are independent.

Definition A.19

Two random variables X, Y on a probability space $(\Omega, \mathfrak{A}, \mathbf{P})$ are said to be independent if the σ-algebras $\sigma(X)$ and $\sigma(Y)$ are independent.
A family $\{X_\alpha\}_\alpha$ of random variables is said to be pairwise independent if X_α and X_β are independent for all $\alpha \neq \beta$.

Independence of random variables X and Y can be conveniently determined using their joint distribution $\mathbf{P}_{X,Y}$: X and Y are independent if and only if $\mathbf{P}_{X,Y}$ equals the product measure $\mathbf{P}_X \times \mathbf{P}_Y$. If X and Y are real-valued with densities p_X and p_Y, they are independent if and only if their joint pdf is

$$p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

Indepenence implies uncorrelatedness

Lemma A.20

If X and Y are independent real-valued random variables and $\mathbf{E}\left[|X|\right], \mathbf{E}\left[|Y|\right] < \infty$, then X and Y are uncorrelated.

Note: The converse is generally false.

Theorem A.21 (Jensen's inequality)

If X is a real-valued random variable with $\mathbf{E}\left[|X|\right] < \infty$ and $\phi : \mathbb{R} \to \mathbb{R}$ a convex function, then

$$\phi(\mathbf{E}\left[X\right]) \leqslant \mathbf{E}\left[\phi(X)\right]. \tag{A.2}$$

Definition A.22

Let $(\Omega, \mathfrak{A}, \mathbf{P})$ be a probability space and let W be a separable Banach space with norm $\|\cdot\|$. We denote by $L^p(\Omega; W)$, $1 \leqslant p < \infty$, the space of W-valued $\mathfrak{A}$-measurable random variables $X : \Omega \to W$ with $\mathbf{E}\left[\|X\|^p\right] < \infty$. The resulting space is a Banach space with the norm

$$\|X\|_{L^p(\Omega;W)} := \left(\int_\Omega \|X(\omega)\|^p \,\mathrm{d}\mathbf{P}(\omega)\right)^{1/p} = \mathbf{E}\left[\|X\|^p\right]^{1/p}.$$

Similarly, $L^\infty(\Omega; W)$ is the Banach space of W-valued random variables $X : \Omega \to W$ for which

$$\|X\|_{L^\infty(\Omega;W)} = \operatorname*{ess\,sup}_{\omega \in \Omega} \|X(\omega)\| < \infty.$$

Probability Theory

Bochner spaces, $p = 2$

The case $p = 2$ when W is a Hilbert space $W = H$ with inner product $(\cdot, \cdot)$ occurs frequently. In this case $L^2(\Omega; H)$ is a Hilbert space with inner product

$$(X, Y)_{L^2(\Omega;H)} := \mathbf{E}\left[(X, Y)\right] = \int_\Omega (X(\omega), Y(\omega))\, \mathrm{d}\mathbf{P}(\omega).$$

Random variables in $L^2(\Omega; H)$ are called mean-square integrable random variables.

For random variables $X, Y \in L^2(\Omega; H)$ the Cauchy-Schwarz inequality takes on the form

$$|(X, Y)_{L^2(\Omega;H)}| \leqslant \|X\|_{L^2(\Omega;H)} \|Y\|_{L^2(\Omega;H)}$$

or

$$\mathbf{E}\left[(X, Y)\right] \leqslant \mathbf{E}\left[\|X\|^2\right]^{1/2} \mathbf{E}\left[\|Y\|^2\right]^{1/2}.$$

Definition A.23

Let H be a separable Hilbert space. A linear operator $C : H \to H$ is the covariance of two H-valued random variables X and Y if

$$(C\phi, \psi) = \mathbf{Cov}((\phi, X), (\psi, Y)) \qquad \forall \phi, \psi \in H.$$

X and Y are said to be uncorrelated if C is the zero operator. If $Y = X$ then C is called the covariance of X.

More generally, the covariance of two random variables X and Y with values in a separable Banach space W may be defined as a bilinear map $c : W' \times W' \to \mathbb{R}$ on the dual space W' of W such that

$$c(\phi, \psi) = \mathbf{Cov}(\langle \phi, X \rangle_{W' \times W}, \langle \psi, Y \rangle_{W' \times W}) \qquad \forall \phi, \psi \in W'.$$

Here $\langle \cdot, \cdot \rangle_{W' \times W}$ denotes the duality bracket between W' and W. The bilinear map c may be identified with a linear operator from $C : W' \to W''$ via the identity

$$\langle C\phi, \psi \rangle_{W'' \times W'} = c(\phi, \psi).$$

Definition A.24

Let W be a Banach space with norm $\|\cdot\|$ and $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of W-valued random variables. We say X_n converges to $X \in W$

almost surely if $X_n(\omega) \to X(\omega)$ for almost all $\omega \in \Omega$, i.e., if

$$\mathbf{P}\left(\|X_n - X\| \to 0 \text{ for } n \to \infty\right) = 1.$$

in probability if $\mathbf{P}\left(\|X_n - X\| > \epsilon\right) \to 0$ for $n \to \infty$ for any $\epsilon > 0$.

in p-th mean or in $L^p(\Omega; W)$ if $\mathbf{E}\left[\|X_n - X\|^p\right] \to 0$ as $n \to \infty$. When $p = 2$ this is known as convergence in mean square.

in distribution if $\mathbf{E}\left[\phi(X_n)\right] \to \mathbf{E}\left[\phi(X)\right]$ as $n \to \infty$ for any bounded and continuous function $\phi : W \to \mathbb{R}$.

Theorem A.25

Let $X_k \to X$ in p-th mean and, for $r > 0$ and a constant $K = K(p)$, assume that

$$\|X_k - X\|_{L^p(\Omega;W)} := \mathbf{E}\left[\|X_k - X\|^p\right]^{1/p} \leqslant \frac{K(p)}{k^r}. \tag{A.3}$$

Then the following convergence properties apply:

(a) $X_k \to X$ in probability and, for any $\epsilon > 0$,

$$\mathbf{P}\left(\|X_k - X\| \geqslant k^{-r+\epsilon}\right) \leqslant \frac{K(p)^p}{k^{p\epsilon}}. \tag{A.4}$$

(b) $\mathbf{E}\left[\phi(X_k)\right] \to \mathbf{E}\left[\phi(X)\right]$ for all Lipschitz continuous functions on W and, if L denotes a Lipschitz constant of ϕ,

$$\left|\mathbf{E}\left[\phi(X_k)\right] - \mathbf{E}\left[\phi(X)\right]\right| \leqslant L\frac{K(p)}{k^r}.$$

(c) If (A.3) holds for all p sufficiently large, then $X_k \to X$ a.s. Furthermore, for each $\epsilon > 0$ there exists a nonnegative random variable K such that $\|X_k(\omega) - X(\omega)\| \leqslant K(\omega)k^{-r+\epsilon}$ for almost all ω.

Contents

Probability Theory

Random vectors

Random variables $\boldsymbol{X} = (X_1, \dots, X_n)^\mathsf{T}$ from $(\Omega, \mathfrak{A}, \mathbf{P})$ to $(\Gamma, \mathfrak{B}(\Gamma)$ with $\Gamma \subset \mathbb{R}^n$ are known as random vectors or multivariate random variables (bivariate for $n = 2$).

Their expected value

$$\boldsymbol{\mu} = \mathbf{E}\left[\boldsymbol{X}\right] = \int_\Omega \boldsymbol{X}(\omega)\,\mathrm{d}\mathbf{P}(\omega) = \left[\mathbf{E}\left[X_1\right], \dots, \mathbf{E}\left[X_n\right]\right]^\mathsf{T}$$

is a vector in $\mathbb{R}^n$. If $\boldsymbol{X}$ has a pdf p, then for $B \in \mathfrak{B}(\Gamma)$

$$\mathbf{P}(\boldsymbol{X} \in B) = \mathbf{P}(\{\omega \in \Omega : \boldsymbol{X}(\omega) \in B\}) = \mathbf{P}_{\boldsymbol{X}}(B) = \int_B p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

The components $\{X_j\}_{j=1}^n$ of $\boldsymbol{X}$ are (pairwise) independent if and only if $\mathbf{P}_{\boldsymbol{X}}$ is the product measure $\mathbf{P}_{X_1} \times \cdots \times \mathbf{P}_{X_n}$. In terms of the pdf, this is equivalent to

$$p(\boldsymbol{x}) = p_{X_1}(x_1) \cdot p_{X_2}(x_2) \cdots p_{X_n}(x_n).$$

Probability Theory

Multivariate uniform

A random vector $\boldsymbol{X} : \Omega \to \Gamma$ with values in a set $\Gamma \subset \mathbb{R}^n$ with finite Lebesgue measure $|\Gamma|$ follows a multivariate uniform distribution on Γ, denoted by

$$\boldsymbol{X} \sim \mathrm{U}(\Gamma)$$

if it has the pdf

$$p(\boldsymbol{x}) \equiv \frac{1}{|\Gamma|}, \qquad \boldsymbol{x} \in \Gamma.$$

Definition A.26

The covariance of two real-valued random vectors $\boldsymbol{X} = [X_1, \dots, X_m]^\mathsf{T}$ and $\boldsymbol{Y} = [Y_1, \dots, Y_n]^\mathsf{T}$ is given by the $m \times n$ matrix

$$\mathbf{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbf{E}\left[(\boldsymbol{X} - \mathbf{E}[\boldsymbol{X}])(\boldsymbol{Y} - \mathbf{E}[\boldsymbol{Y}])^\mathsf{T}\right].$$

$\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to be uncorrelated if $\mathbf{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{O}$ (the $m \times n$ zero matrix). The matrix $\mathbf{Cov}(\boldsymbol{X}, \boldsymbol{X}) \in \mathbb{R}^{n \times n}$ is called the covariance matrix of $\boldsymbol{X}$.

Proposition A.27

Let $\boldsymbol{X}$ be an $\mathbb{R}^n$-valued random variable with mean vector $\boldsymbol{\mu}$ and covariance matric $\boldsymbol{C}$. Then $\boldsymbol{C}$ ist symmetric positive semi-definite and its trace is given by $\mathbf{E}\left[\|\boldsymbol{X} - \boldsymbol{\mu}\|_2^2\right]$.

A random vector with mean vector $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{C}$ is said to follow an n-variate Gaussian distribution if it has the pdf

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d \det \boldsymbol{C}}} \exp\left(\frac{-(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T} \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right). \tag{A.5}$$

To cover the case that $\boldsymbol{C}$ is singular we introduce the characteristic function.

Definition A.28

The characteristic function of an $\mathbb{R}^n$-valued random vector $\boldsymbol{X}$ is $\mathbf{E}\left[\exp(i\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{X})\right]$, for $\boldsymbol{\lambda} \in \mathbb{R}^n$. If $\boldsymbol{X}$ has the pdf p, then its characteristic function is

$$\mathbf{E}\left[\exp(i\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{X})\right] = (2\pi)^{n/2}\hat{p}(-\boldsymbol{\lambda}),$$

where $\hat{p}$ is the Fourier transform of p. (The minus sign is a convention in probability theory.)

Proposition A.29

A random vector $\boldsymbol{X}$ has the density (A.5) for a given vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and symmetric positive definite matrix $\boldsymbol{C} \in \mathbb{R}^{n \times n}$ if and only if its characteristic function is

$$\mathbf{E}\left[\exp(i\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{X})\right] = \exp(i\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{C}\boldsymbol{\lambda}). \tag{A.6}$$

Definition A.30

An $\mathbb{R}^n$-valued random vector $\boldsymbol{X}$ follows a multivariate normal (or Gaussian) distribution, denoted

$$\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{C}),$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{C} \in \mathbb{R}^{n \times n}$ is symmetric positive semi-definite, if its characteristic function is (A.6).

If $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{C})$ is a multivariate normal random vector, then for any $\boldsymbol{a} \in \mathbb{R}^n$ the linear combination

$$Y = \boldsymbol{a}^\top \boldsymbol{X} = \sum_{k=1}^{n} a_k X_k$$

follows the normal distribution $Y \sim \mathrm{N}(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \boldsymbol{C} \boldsymbol{a})$.

Contents

5 Probability Theory

5.1 Random Variables

5.2 Random vectors

5.3 Limit Theorems

5.4 Statistical Estimation

6 Elliptic Boundary Value Problems

7 Collection of Results from Functional Analysis

8 Miscellanea

Definition A.31

A sequence $\{X_j\}_{j\in\mathbb{N}}$ of random variables is said to be independent and identically distributed (i.i.d.) if they all follow the same probability distribution and, in addition, are pairwise independent.

The classical limit theorems of probability theory concern sums of iid random variables. For an iid sequence $\{X_j\}_{j\in\mathbb{N}}$, we introduce the notation

$$S_n := X_1 + \cdots + X_n, \qquad n \in \mathbb{N}.$$

Theorem A.32 (Chebyshev inequality)

A random variable X with finite mean μ and finite variance σ^2 satisfies

$$c^2\mathbf{P}(|X-\mu| \geqslant c) \leqslant \sigma^2.$$

Theorem A.33 (WLLN)

Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of i.i.d. random variables on a given probability space $(\Omega, \mathfrak{A}, \mathbf{P})$ with mean μ and finite variance. Then

$$\frac{S_n}{n} \to \mu \qquad \text{in probability, i.e.}$$

for ever fixed $\epsilon > 0$ there holds

$$\mathbf{P}\left(|S_n/n - \mu| > \epsilon\right) \to 0 \qquad \text{as} \qquad n \to \infty.$$

Theorem A.34 (SLLN)

Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of i.i.d. real-valued random variables on a given probability space $(\Omega, \mathfrak{A}, \mathbf{P})$. Then S_n/n has a finite limit if and only if $\mathbf{E}\left[|X_1|\right] < \infty$, in which case

$$\frac{S_n}{n} \to \mathbf{E}\left[X_1\right] \qquad \text{a.s.}$$

If $\mathbf{E}\left[|X_1|\right] = \infty$, then $\limsup_{n\to\infty} |S_n|/n \to \infty$ a.s.

Lemma A.35 (Kronecker's Lemma)

If the series $\sum_{k=1}^{\infty} x_k/k$ converges (not necessarily absolutely) for a sequence $\{x_k\}_{k\in\mathbb{N}}$ of real numbers, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} x_k = 0.$$

Lemma A.36

The sequence $\{X_k\}_{k\in\mathbb{N}}$ converges a.s. if and only if

$$\lim_{n\to\infty} \mathbf{P}\{\sup_{k\in\mathbb{N}} |X_{n+k} - X_n| > \epsilon\} = 0 \qquad \forall \epsilon > 0.$$

Theorem A.37 (Kolmogorov Inequality)

Let $X_1, \dots, X_n$ be independent real-valued random variables with $\mathbf{E}[X_j] = 0$ and $0 < \sigma_j^2 = \mathbf{Var}\, X_j < \infty$ for all j. Then for each $\epsilon > 0$

$$\mathbf{P}\left\{\max_{1 \leqslant k \leqslant n} |S_k| > \epsilon\right\} \leqslant \frac{1}{\epsilon^2} \sum_{j=1}^{n} \sigma_j^2. \tag{A.7}$$

Conversely, if there exists c such that $\mathbf{P}\{|X_k| < \epsilon\} = 1$ for each k, then for each ϵ

$$\mathbf{P}\left\{\max_{1 \leqslant k \leqslant n} |S_k| > \epsilon\right\} \geqslant 1 - \frac{(c+\epsilon)^2}{\sum_{j=1}^{n} \sigma_j^2}. \tag{A.8}$$

Theorem A.38

Let $\{X_k\}_{k \in \mathbb{N}}$ be independent real-valued random variables with $\mathbf{E}[X_k] = 0$ for all k. If

$$\sum_{k=1}^{\infty} \mathbf{E}\left[X_k^2\right] = \sum_{k=1}^{\infty} \mathbf{Var}\, X_k < \infty$$

then $\sum_{k=1}^{\infty} X_k$ converges a.s.

Definition A.39

For a real-valued random variable X and $c > 0$ we denote the truncation of X at c by

$$X^c := X\mathbb{1}_{\{|X| \leqslant c\}} = \begin{cases} X & \text{if } |X| \leqslant c, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem A.40 (Three-series theorem)

Let $\{X_k\}_{k\in\mathbb{N}}$ be independent. If, for some $c > 0$,

$$\sum_{k=1}^{\infty} \mathbf{P}\{|X_k| > c\} < \infty, \tag{A.9a}$$

$$\sum_{k=1}^{\infty} |\mathbf{E}\,[X_k^c]\,| < \infty, \tag{A.9b}$$

$$\sum_{k=1}^{\infty} \mathbf{Var}\, X_k^c < \infty, \tag{A.9c}$$

then $\sum_{k=1}^{\infty} X_k$ converges a.s.
Conversely, if $\sum_{k=1}^{\infty} X_k$ converges a.s., then (A.9a)–(A.9c) hold for every $c > 0$.

Let the sequence $\{X_j\}_{j\in\mathbb{N}}$ of real-valued random variables be independent, but not necessarily identically distributed. In addition, let $\mathbf{E}[X_j] = 0$ and $\mathbf{E}\left[X_j^2\right] < \infty$ for all j.

Besides $S_n = \sum_{j=1}^n X_j$, introduce the quantities

$$\sigma_j^2 := \mathbf{Var}\, X_j,$$
$$\Sigma_n^2 := \sum_{j=1}^n \sigma_j^2 = \mathbf{Var}\, S_n.$$

The central limit theorem (CLT) is the statement that

$$\lim_{n\to\infty} \frac{S_n}{\Sigma_n} = \lim_{n\to\infty} \frac{S_n - \mathbf{E}[S_n]}{\sqrt{\mathbf{Var}\, S_n}} \sim \mathrm{N}(0,1) \quad \text{in distribution.}$$

Definition A.41 (Lyapunov condition)

The sequence $\{X_k\}_{k\in\mathbb{N}}$ satisfies the Lyapunov condition if $\mathbf{E}\left[|X_k|^3\right] < \infty$ for each k and

$$\lim_{n\to\infty} \frac{1}{\Sigma_n^2} \sum_{k=1}^{n} \mathbf{E}\left[|X_k|^3\right] = 0.$$

Theorem A.42 (CLT)

If $\{X_k\}_{k\in\mathbb{N}}$ satisfies the Lyapunov condition, then $S_n/\Sigma_n \to \mathrm{N}(0,1)$ in distribution.

Definition A.43 (Lindeberg condition)

The sequence $\{X_k\}_{k\in\mathbb{N}}$ satisfies the Lindeberg condition if for every $\epsilon > 0$

$$\lim_{n\to\infty} \frac{1}{\Sigma_n^2} \sum_{k=1}^{n} \mathbf{E}\left[X_k^2 \cdot \mathbb{1}_{\{|X_k|>\epsilon\Sigma_n\}}\right] = 0.$$

Proposition A.44

The Lyapunov condition implies the Lindeberg condition.

Example A.45

(1) If $\mathbf{P}\{|X_k| \leqslant c\} = 1$ for some constant c and if $\Sigma_n^2 \to \infty$, then the Lindeberg condition is satisfied.

(2) If $\{X_k\}_{k\in\mathbb{N}}$ are i.i.d. with variance $\sigma^2 \in (0,\infty)$, then the Lindeberg condition is satisfied.

Theorem A.46 (Lindeberg-Feller CLT)

If $\{X_k\}_{k\in\mathbb{N}}$ satisfies the Lindeberg condition, then $S_n/\Sigma_n \to \mathrm{N}(0,1)$ in distribution.

Theorem A.47 (Berry, 1941; Esseen 1942)

Let $\{X_k\}_{k\in\mathbb{N}}$ be i.i.d. random variables with (common)

$$\mu := \mathbf{E}\left[X_1\right], \quad \sigma^2 := \mathbf{Var}\, X_1 > 0, \quad \rho := \mathbf{E}\left[|X_1 - \mu|^3\right] < \infty.$$

If F_n denotes the distribution function of $(S_n - n\mu)/(\sigma\sqrt{n})$ and Φ that of the standard normal distribution $\mathrm{N}(0,1)$, then, with a universal constant C,

$$\sup_{x\in\mathbb{R}} |\Phi(x) - F_n(x)| \leqslant C \cdot \frac{\rho}{\sigma^3\sqrt{n}}.$$

Note: the constant C is known to satisfy $0.4097 \leqslant C \leqslant 0.7056$ [Shevtsova, 2007].

Contents

5 Probability Theory

5.1 Random Variables

5.2 Random vectors

5.3 Limit Theorems

5.4 Statistical Estimation

6 Elliptic Boundary Value Problems

7 Collection of Results from Functional Analysis

8 Miscellanea

Statistical Estimation

- Estimation theory is concerned with determining an unknown quantity θ associated with the probability distribution of a random variable X given n i.i.d. samples $\{X_k\}_{k=1}^n$ of X.
- Typical examples of such quantities θ are moments of X's distribution such as the mean and the variance. Another common situation is the estimation of one or more parameters which determine the distribution of X.
- An estimator for a scalar quantity θ is a function

  $$\phi : \mathbb{R}^n \to \mathbb{R}, \qquad \hat{\theta} = \phi(X_1, \dots, X_n)$$

  mapping n i.i.d. realizations of X to the estimate $\hat{\theta}$ of θ.
- Note that, since each of the n random samples X_k are random variables, the same is true of

  $$\hat{\theta} = \hat{\theta}(\omega) = \phi(X_1(\omega), \dots, X_n(\omega)).$$

  Once the samples have been drawn/realized, the estimate $\hat{\theta}$ is a real number.

Statistical Estimation

Sample average, unbiased estimator

- The sample average

$$\hat{\mu}_n := \frac{X_1 + \cdots + X_n}{n}$$

  is an estimate for the mean $\mu = \mathbf{E}[X]$.

- Since the X_k are i.i.d., we conclude from the linearity of expectation that

$$\mathbf{E}[\hat{\mu}_n] = \frac{1}{n} \sum_{k=1}^{n} \mathbf{E}[X_k] = \frac{1}{n} \cdot n\mu = \mu.$$

- If $\mathbf{E}[|X|] < \infty$ the SLLN tells us that also $\hat{\mu}_n \to \mu = \mathbf{E}[X]$ a.s. as $n \to \infty$.
- Since $\mathbf{Var}\, \hat{\mu}_n = \frac{\sigma^2}{n}$, where $\sigma^2 = \mathbf{Var}\, X$, we note that the variance $\hat{\mu}_n$ decreases like $1/n$ with growing sample size.

Definition A.48

An estimator for which $\mathbf{E}\left[\hat{\theta}\right] = \theta$ is called unbiased.

Statistical Estimation

Sample variance

The sample variance

$$\hat{\sigma}_n^2 := \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \hat{\mu}_n)^2$$

is an unbiased estimator for $\sigma^2 = \textbf{Var}\, X$.

In addition, there holds $\hat{\sigma}_n^2 \to \sigma^2$ a.s. as $n \to \infty$.

Statistical Estimation

Confidence intervals

An estimator $\hat{\theta}$ is, in general, only close to the estimated quantity θ in a probabilistic sense, i.e., it will fluctuate around the true value θ from realization to realization.

For a probability distribution depending on a real-valued parameter θ, we denote by

$$\mathbf{P}(A \,|\, \theta)$$

the probability of event A if the true value of the parameter is θ.

Definition A.49

Given n i.i.d. random variables $\{X_k(\omega)\}_{k=1}^n$ and a number $\gamma \in [0,1]$, a confidence interval of level γ for a quantity θ is determined by two functions $\tau_-, \tau^+ : \mathbb{R}^n \to \mathbb{R}$ such that, for all possible values of θ,

$$\mathbf{P}\left(\tau_-(X_1,\dots,X_n) \leqslant \theta \leqslant \tau_+(X_1,\dots,X_n) \,|\, \theta\right) = \gamma.$$

Confidence intervals example

As an example, take the random variables

$$X_k = \mu + \epsilon_k, \qquad \mu \in \mathbb{R}, \quad \epsilon_k \sim \mathrm{N}(0,1) \text{ i.i.d.}, \quad k = 1, \dots, n.$$

Then $\mu = \mathbf{E}[X]$ and for the estimation error we obtain

$$\hat{\mu}_n - \mu = \frac{1}{n} \sum_{k=1}^{n} \epsilon_k \sim \mathrm{N}(0, \tfrac{1}{n}).$$

and therefore $\sqrt{n}(\hat{\mu}_n - \mu) \sim \mathrm{N}(0,1)$.

Given $\gamma \in [0,1]$ we choose $a \geqslant 0$ such that $\Phi(a) - \Phi(-a) = \gamma$ and obtain

$$\gamma = \mathbf{P}(-a \leqslant \sqrt{n}(\hat{\mu}_n - \mu) \leqslant a \,|\, \mu) = \mathbf{P}\left(\hat{\mu}_n - \frac{a}{\sqrt{n}} \leqslant \mu \leqslant \hat{\mu}_n + \frac{a}{\sqrt{n}} \,|\, \mu\right),$$

so that $\tau_\pm = \hat{\mu}_n \pm \frac{a}{\sqrt{n}}$ yield a confidence interval of level γ for μ.

Contents

5 Probability Theory

6 Elliptic Boundary Value Problems

7 Collection of Results from Functional Analysis

8 Miscellanea

Contents

Elliptic Boundary Value Problem

We consider the elliptic boundary value problem of finding the solution of the partial differential equation with Dirichlet boundary condition

$$-\nabla\cdot(a\nabla u) = f \qquad \text{on } D \subset \mathbb{R}^2, \tag{B.1a}$$

$$u = g \qquad \text{on } \partial D, \tag{B.1b}$$

given a convex bounded domain D with sufficiently smooth boundary ∂D, a coefficient function $a : D \to \mathbb{R}^+$, a source term $f : D \to \mathbb{R}$ and boundary data in the form of a function $g : \partial D \to \mathbb{R}$.

The differential operator in (B.1a) is short for

$$\nabla\cdot(a\nabla u) = \sum_{j=1}^{2} \frac{\partial}{\partial x_j}\left(a(\boldsymbol{x})\frac{\partial u(\boldsymbol{x})}{\partial x_j}\right)$$

Equation (B.1a) is a model for diffusion phenomena occurring in , e.g., heat conduction, electrostatics, potential flow and elasticity. Generalizations of (B.1) involve the addition of lower-order terms, other boundary conditions, a matrix-valued coefficient function and dependence of a on u.

Elliptic Boundary Value Problem

Strong and weak solution

If $f \in C(\overline{D})$ and $a \in C^1(\overline{D})$, then a function $u \in C^2(D) \cap C^1(\overline{D})$ which satisfies (B.1) is called a classical solution or a strong solution of the boundary value problem.

There are (theoretical and practical) reasons for generalizing the classical solution concept. The key to this generalization lies in reformulating (B.1) as a variational problem. Multiplying both sides of (B.1a) by an arbitrary function $\phi \in C_0^\infty(D)$, in this context known as a test function, and integrating by parts, we observe that any (classical) solution of (B.1) also satisfies the equation

$$a(u,\phi) = \ell(\phi) \qquad \forall \phi \in C_0^\infty(D), \tag{B.2}$$

with the symmetric bilinear form $a(\cdot,\cdot)$ and linear functional $\ell(\cdot)$ given by

$$a(u,\phi) = \int_D a(\boldsymbol{x}) \nabla u(x) \cdot \nabla \phi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \qquad \ell(\phi) = \int_D f(\boldsymbol{x}) \phi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{B.3}$$

For (B.2) to make sense, it is sufficient that the integrals and derivatives are well-defined.

Elliptic Boundary Value Problem

Strong and weak solution

This is the case if u and ϕ are taken to lie in the Sobolev space

$$H^1(D) := \{v \in L^2(D) : \nabla v \in L^2(D)^2\},$$

which is a Hilbert space with respect to the inner product

$$(u, v)_{H^1(D)} = \int_D (\nabla u \cdot \nabla v + uv)\, \mathrm{d}\boldsymbol{x} = (\nabla u, \nabla v) + (u, v),$$

where we use $(\cdot, \cdot)$ to denote the inner product in $L^2(D)$. The associated norm on $H^1(D)$ is

$$\|u\|^2_{H^1(D)} = \int_D \left(|\nabla u|^2 + u^2\right)\, \mathrm{d}\boldsymbol{x}.$$

The gradients are in terms of weak derivatives in the sense

$$\left(\frac{\partial u}{\partial x_j}, \phi\right) = -\left(u, \frac{\partial \phi}{\partial x_j}\right) \qquad \forall \phi \in C_0^\infty(D).$$

Elliptic Boundary Value Problem

Strong and weak solution

Stating the boundary condition (B.1b) requires a well-defined notion of evaluating a function from $H^1(D)$ on the lower-dimensional manifold ∂D.

- Functions in $H^1(D)$ satisfying the BC with homogeneous boundary data $g \equiv 0$ are easily defined as lying in the subspace

$$H_0^1(D) := \overline{C_0^\infty(D)}^{\|\cdot\|_{H_1(D)}} \subset H^1(D).$$

- For inhomogeneous boundary data we define the space

$$W := H_g^1(D) := \{v \in H^1(D) : u_{|\partial D} = g\}.$$

The evaluation on the boundary is understood in the following sense: for a sufficiently smooth boundary there exists a bounded trace operator $\gamma : H^1(D) \to L^2(\partial D)$ such that for all $u \in C^1(\overline{D})$ there holds $\gamma u = u_{|\partial D}$. Since $C^1(\overline{D})$ is dense in $H^1(D)$, we have $\gamma u = \lim_{n\to\infty} u_{|\partial D}$ for any approximating sequence $\{u_n\} \subset C^1(\overline{D})$ converging to u in $H^1(D)$.

Elliptic Boundary Value Problem

Strong and weak solution

Definition B.1

The trace space of $H^1(D)$ for a sufficiently smooth domain D is defined as

$$H^{1/2}(\partial D) := \gamma(H^1(D)) = \{\gamma u : u \in H^1(D)\}.$$

$H^{1/2}(\partial D)$ is a Hilbert space with norm

$$\|g\|_{H^{1/2}(\partial D)} := \inf\{\|u\|_{H^1(D)} : \gamma u = g, u \in H^1(D)\}.$$

Sine in general $H^{1/2}(\partial D) \subsetneq L^2(\partial D)$, boundary data g in (B.1b) must be chosen from $H^{1/2}(\partial D)$.

Lemma B.2

There exists $K_\gamma > 0$ such that, for all $g \in H^{1/2}(\partial D)$, we can find $u_g \in H^1(D)$ with $\gamma u_g = g$ and

$$\|u_g\|_{H^1(D)} \leqslant K_\gamma \|g\|_{H^{1/2}(\partial D)}$$

Elliptic Boundary Value Problem

Strong and weak solution

We denote the spaces of trial and test functions by

$$W := H^1_g(D), \qquad \text{and} \qquad V := H^1_0(D).$$

Assumption B.3

The coefficient function $a = a(\boldsymbol{x})$ in (B.1a) satisfies

$$0 < a_{\min} \leqslant a(\boldsymbol{x}) \leqslant a_{\max} < \infty \qquad \text{for almost all } \boldsymbol{x} \in D$$

for positive constants $a_{\min}$ and $a_{\max}$. In particular, $a \in L^\infty(D)$ and a is uniformly bounded away from zero.

By Assumption B.3, the bilinear form $a(\cdot,\cdot)$ is bounded on $H^1(D)$, i.e.,

$$|a(u,v)| \leqslant C\|u\|_{H^1(D)}\|v\|_{H^1(D)}, \qquad \forall u, v \in H^1(D)$$

with a constant $C \leqslant \|a\|_{L^\infty(D)}$.

Elliptic Boundary Value Problem

Strong and weak solution

Definition B.4

A weak solution of (B.1) is a function $u \in W$ such that

$$a(u,v) = \ell(v) \qquad \forall v \in V, \tag{B.4}$$

with $a(\cdot,\cdot)$ and $\ell(\cdot)$ as defined in (B.3).

Definition B.5

A bilinear form $a : H \times H \to \mathbb{R}$ on a Hilbert space H is said to be coercive if there exists a constant $\alpha > 0$ such that

$$a(u,u) \geqslant \alpha \|u\|_H^2 \qquad \forall u \in H.$$

Lemma B.6 (Lax & Milgram)

Let H be a real Hilbert space with norm $\|\cdot\|$ and let ℓ be a bounded linear functional on H. Let $a : H \times H \to \mathbb{R}$ be a bilinear form that is bounded and coercive. Then there exists a unique $u_\ell \in H$ such that $a(u_\ell, v) = \ell(v)$ for all $v \in H$.

Elliptic Boundary Value Problem

Strong and weak solution

For functions in $H^1(D)$ we introduce the H^1 semi-norm

$$|u|_{H^1(D)} := \left(\int_D |\nabla u|^2 \,\mathrm{d}\boldsymbol{x}\right)^{1/2}.$$

as well as the energy norm associated with the coefficient function a as

$$|u|_a := a(u,u)^{1/2} = \left(\int_D a\nabla u \cdot \nabla u \,\mathrm{d}\boldsymbol{x}\right)^{1/2}.$$

Theorem B.7 (Poincaré-Friedrichs inequality)

For a bounded domain D there exists a constant $C = C_D > 0$ such that

$$\|u\|_{L^2(D)} \leqslant C_D |u|_{H^1(D)} \qquad \forall u \in H_0^1(D).$$

Elliptic Boundary Value Problem

Strong and weak solution

Lemma B.8

Under Assumption B.3 the bilinear form $a : H^1(D) \times H^1_0(D) \to \mathbb{R}$ is bounded and the energy norm is equivalent to the H^1 semi-norm on $H^1(D)$.

Theorem B.9

Let Assumption B.3 hold, $f \in L^2(D)$ and $g \in H^{1/2}(\partial D)$. Then (B.1) has a unique weak solution $u \in W = H^1_g(D)$.

Theorem B.10

Under the conditions of Theorem B.9 the weak solution $u \in W$ satisfies

$$|u|_{H^1(D)} \leqslant K \left(\|f\|_{L^2(D)} + \|g\|_{H^{1/2}(\partial D)} \right)$$

where $K = \max\{C_D/a_{\min}, K_\gamma(1 + a_{\max}/a_{\min})\}$.

Elliptic Boundary Value Problem

Perturbed data

Replacing a und f in (B.1) by approximations $\tilde{a}$ and $\tilde{f}$, leads to the perturbed problem of finding $\tilde{u} \in W$ such that

$$\tilde{a}(\tilde{u}, v) = \tilde{\ell}(v) \qquad \forall v \in V \tag{B.5}$$

with $\tilde{a} : W \times V \to \mathbb{R}$ sowie $\tilde{\ell} : V \to \mathbb{R}$ defined by

$$\tilde{a}(u, v) = \int_D \tilde{a}(\boldsymbol{x}) \nabla u(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \qquad \tilde{\ell}(\phi) = \int_D \tilde{f}(\boldsymbol{x}) v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{B.6}$$

Theorem B.11

Let Assumption B.3 hold for a as well as for $\tilde{a}$ with constants $\tilde{a}_{\min}$, $\tilde{a}_{\max}$. If, furthermore, $\tilde{f} \in L^2(D)$ and $g \in H^{1/2}(\partial D)$, then problem (B.5) has a unique weak solution $\tilde{u} \in W = H^1_g(D)$.

Elliptic Boundary Value Problem

Perturbed data

Theorem B.12

Under the conditions of Theorems B.9 and B.11, if $u, \tilde{u} \in W$ denote the solutions of (B.4) and (B.5), respectively, then

$$|u - \tilde{u}|_{H^1(D)} \leqslant C_D \tilde{a}_{\min}^{-1} \|f - \tilde{f}\|_{L^2(D)} + \tilde{a}_{\min}^{-1} \|a - \tilde{a}\|_{L^\infty(D)} |u|_{H^1(D)}$$

Contents

5. Probability Theory

6. Elliptic Boundary Value Problems
   - 6.1 Weak Formulation
   - 6.2 Finite Element Approximation
   - 6.3 Finite Element Convergence

7. Collection of Results from Functional Analysis

8. Miscellanea

Finite Element Approximation

Galerkin discretization

Given: linear variational problem of finding $u \in V$, V a Hilbert space with norm $\|\cdot\|$, such that

$$a(u, v) = \ell(v) \qquad \forall v \in V \tag{B.7}$$

with a bilinear form $a(\cdot,\cdot)$ and linear form $\ell(\cdot)$ on V which satisfy the assumptions of the Lax-Milgram lemma.

Galerkin method for finding approximate solutions of (B.7) proceeds by restricting the problem to a finite-dimensional subspace $V_n \subset V$: denote by $u_n \in V_n$ the solution of

$$a(u_n, v) = \ell(v) \qquad \forall v \in V_n. \tag{B.8}$$

Note: The Galerkin approximation u_n of u with respect to the space V_n is uniquely determined since the conditions of the Lax-Milgram lemma are satisfied for Problem (B.8) by inclusion.

Finite Element Approximation

Céa's lemma

The simple structure of a linear variational problem allows its reduction to a problem of best approximation.

Lemma B.13 (Céa)

If the assumptions of the Lax-Milgram lemma apply to Problem (B.7) with solution $u \in V$, then the Galerkin approximation u_n, i.e., the solution of (B.8), satisfies

$$\|u - u_n\| \leqslant \frac{C}{\alpha} \inf_{v \in V_n} \|u - v\|. \tag{B.9}$$

Finite Element Approximation

Céa's lemma, symmetric case

- If the bilinear form $a(\cdot,\cdot)$ is, in addition, symmetric (Hermitian) then, because of coercivity, it defines an inner product on V.
- Galerkin orthogonality then implies u_n is the a-orthogonal projection of u onto V_n and therefore the best approximation to u from V_n with respect to the associated (energy) norm.
- In the energy norm (B.9) is therefore satisfied with $C = \alpha = 1$.
- Coercivity and boundedness also imply that the energy norm is equivalent with $\|\cdot\|$, i.e.,

  $$\sqrt{\alpha}\|v\| \leqslant |v|_a \leqslant \sqrt{C}\|v\| \qquad \forall v \in V,$$

  which leads to the improved estimate over (B.9)

  $$\|u - u_n\| \leqslant \sqrt{\frac{C}{\alpha}} \inf_{v \in V_n} \|u - v\|.$$

Finite Element Approximation

Application to elliptic BVP

We have seen that, for the elliptic BVP (B.1), we have the equivalences

$$\|\cdot\|_{H^1(D)} \asymp |\cdot|_{H^1(D)} \asymp |\cdot|_a.$$

Corollary B.14

Under Assumption B.3, the Galerkin approximation u_n fo the solution of the elliptic boundary value problem (B.1), with respect to any subspace V_n of $V = H_0^1(D)$, satisfies

$$|u - u_n|_a = \inf_{v \in V_n} |u - v|_a,$$

$$|u - u_n|_{H^1(D)} \leqslant \sqrt{\frac{a_{\min}}{a_{\max}}}\, |u - v|_{H^1(D)} \qquad \forall v \in V_n.$$

Finite Element Approximation

Galerkin system

Given a basis $\{v_1, \dots, v_n\}$ of V_n and the solution $u_n = \sum_{j=1}^n \xi_j v_j$, then the Galerkin variational equation (B.8) is equivalent with

$$\sum_{j=1}^n \xi_j \, a(v_j, v_i) = \ell(v_i), \qquad i = 1, \dots, n,$$

which, when rewritten as a linear system of equation, becomes the Galerkin system

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{B.10}$$

with Galerkin matrix $[\boldsymbol{A}]_{i,j} = a(v_j, v_i)$, unknown vector $[\boldsymbol{x}]_i = \xi_i$ and right-hand side vector $[\boldsymbol{b}]_i = \ell(v_i)$.

- If $a(\cdot, \cdot)$ is symmetric, then so is $\boldsymbol{A}$.
- If $a(\cdot, \cdot)$ is coercive, then $\boldsymbol{A}$ is (uniformly) positive definite.

Finite Element Approximation

The finite element method

- Different Galerkin methods result from different choices of subspaces.
- Wavelets.
- Trigonometric functions, global polynomials (spectral methods).
- Radial basis functions.
- The finite element method employs finite dimensional subspaces of the variational spaces (trial and test spaces) consisting of piecewise polynomials with respect to a partition of D.
- We shall assume in the following that D is a polygon (polyhedron), but the finite element method can also be applied to domains with curved boundaries.

Finite Element Approximation

Triangulations

Assumptions on the partition of the domain D, denoted by $\mathscr{T}_h$ with elements K:

(Z$_1$) $\overline{D} = \cup_{K\in\mathscr{T}_h} K$.

(Z$_2$) Each $K \in \mathscr{T}_h$ is a closed set with nonempty interor $\mathring{K}$.

(Z$_3$) For two distinct $K_1, K_2 \in \mathscr{T}_h$ there holds $\mathring{K}_1 \cap \mathring{K}_2 = \emptyset$.

(Z$_4$) Each $K \in \mathscr{T}_h$ has a Lipschitz-continuous boundary ∂K.

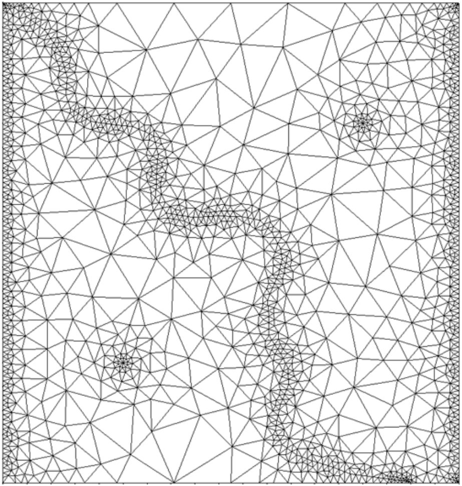
The partition is usually assigned a discretization parameter $h > 0$ given by

$$h := \max_{K\in\mathscr{T}^h} \operatorname{diam} K,$$

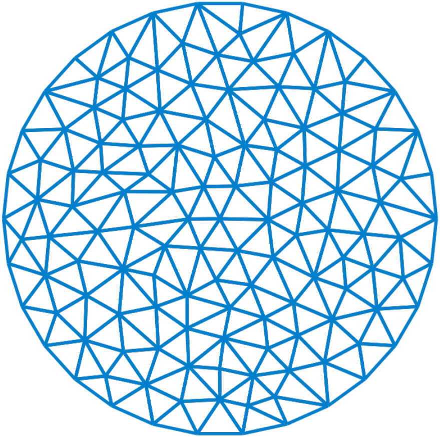
which is a measure of how fine the partition is.

Finite Element Approximation

Triangulations



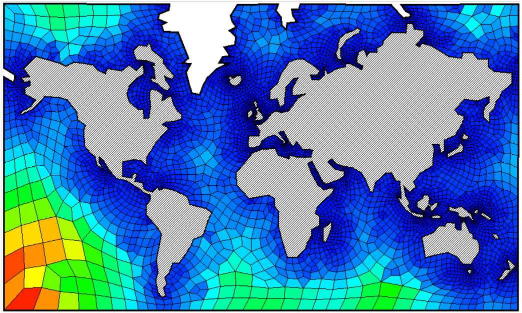
Triangular mesh on a square domain.



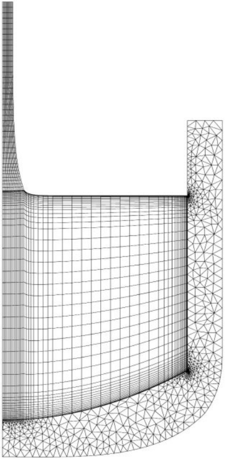
Triangular mesh on a polygonal approximation of a circle.

Finite Element Approximation

Triangulations



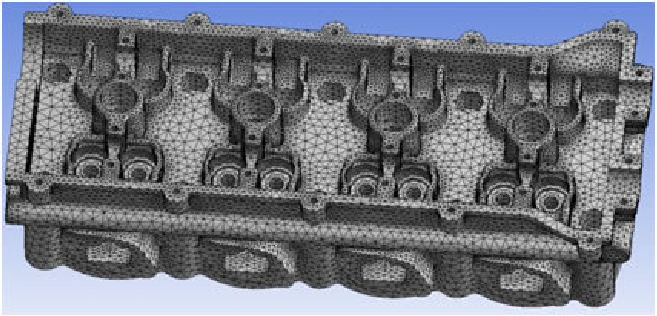
Quadrilateral mesh on a rectangular (exterior) domain.



Mesh consisting of triangles and quadrilaterals.

Finite Element Approximation

Triangulations



Tetrahedral mesh of complex 3D geometry (engine block).

Finite Element Approximation

H^1-conforming finite element spaces

A conforming Galerkin approximation is one which employs finite-dimensional spaces V_n such that $V_n \subset V$.

Let V^h denote a space of piecewise continuous functions $v : \overline{D} \to \mathbb{R}$ with respect to an admissible triangulation $\mathscr{T}_h$ of D, i.e., such that each restriction $v|_K$ to any $K \in \mathscr{T}_h$ is continuous on K.

Theorem B.15

With the notation defined above, there holds $V^h \subset H^1(D)$ if, and only if,

$$V^h \subset C(\overline{D}) \qquad \text{and} \qquad \{v|_K : v \in V^h\} \subset H^1(K).$$

In this case $\{v \in V^h : v = 0 \text{ on } \partial D\} \subset H^1_0(D)$.

Finite Element Approximation

Finite elements

According to [Ciarlet, 1978], a finite element is a triple (K, P_K, Ψ_K) such that

(1) K is a nonempty set

(2) P_K is a finite-dimensional space of functions defined on K and

(3) Ψ_K is a set of linearly independent linear functionals ψ on P_K with the property that, for any $p \in P_K$,

$$\psi(p) = 0 \ \ \forall \psi \in \Psi_K \qquad \Rightarrow \qquad p = 0.$$

We shall consider a single finite element, the so-called linear triangle, where

(1) $K \in \mathbb{R}^2$ is a triangle with (non-collinear) vertices $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$,

(2) P_K is the space of all affine functions on K and

(3) Ψ_K consists of the three functionals

$$\Psi_K = \{\psi_j : P_K \to \mathbb{R}, \psi_j(p) = p(\boldsymbol{x}_j), j = 1, 2, 3\}.$$

Finite Element Approximation

Trianglular finite elements

- To construct a (global) finite element space V^h based on linear triangle elements consider a triangulation $\mathscr{T}^h$ of D consisting of (closed) triangles K which satisfy properties (**Z1**)–(**Z4**).
- The functions in V^h will also lie in $H^1(D)$ if they are continuous on $\overline{D}$, which, for piecewise linear (polynomial) functions, is equivalent with their being continuous across triangle boundaries.
- We thus obtain the space

$$V^h := \{v \in C(\overline{D}) : v|_K \in \mathscr{P}_1 \ \forall K \in \mathscr{T}^h\},$$

  where $\mathscr{P}_k$ denotes the space of (multivariate) polynomials of (complete) degree k.
- A subspace V_0^h of V^h is given by

$$V_0^h := \{v \in V^h : v|_{\partial D} = 0\} \subset H_0^1(D).$$

Finite Element Approximation

Degrees of freedom, nodal basis

- A continuous piecewise linear function in V^h is completely determined by its values at all triangle vertices.
- Such a (finite) set of parameters which uniquely determine a finite element function is called a set of degrees of freedom (DOF).
- In V_0^h these are the values at all nodes which do not lie on ∂D; denote their number by n.
- A particularly convenient basis $\{\phi_1, \dots, \phi_n\}$ of V_0^h is the so-called nodal basis characterized by

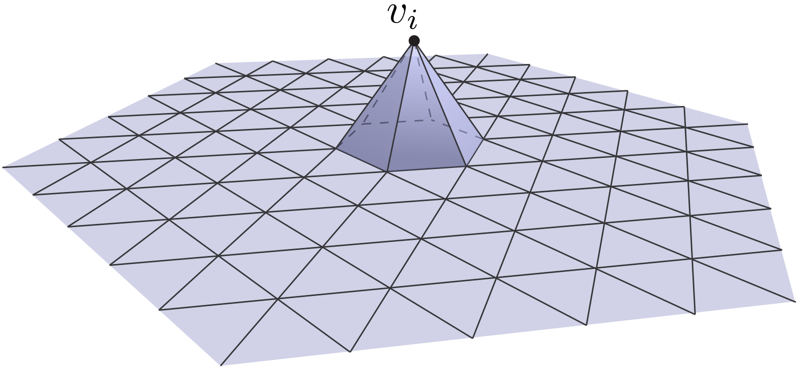
$$\phi_j(\boldsymbol{x}_i) = \delta_{i,j} \qquad i, j = 1, \dots, n.$$

- If $\mathscr{N}^h = \{x_1, \dots, x_n\}$ denotes the set of vertices $x_j \notin \partial D$, then

$$\operatorname{supp} \phi_j = \bigcup_{\substack{K \in \mathscr{T}^h \\ \boldsymbol{x}_j \in K}} K.$$

Finite Element Approximation

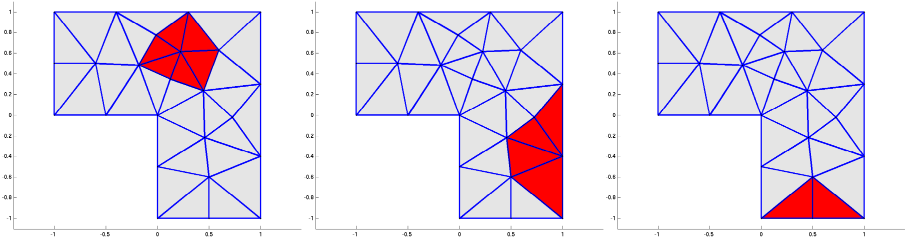
Nodal basis for linear triangles



A nodal basis function with its support.

Finite Element Approximation

Nodal basis for linear triangles



Triangulation of an L-shaped domain with the supports of several basis functions.

Finite Element Approximation

Galerkin matrix, linear triangles

Implications for Galerkin system (B.10):

$$[\boldsymbol{b}]_i = \ell(\phi_i) = \int_D f\phi_i \,\mathrm{d}\boldsymbol{x} = \int_{\operatorname{supp}\phi_i} f\phi_i \,\mathrm{d}\boldsymbol{x},$$

$$\begin{aligned}[\boldsymbol{A}]_{i,j} = a(\phi_j,\phi_i) &= \int_D a(\boldsymbol{x})\phi_i(\boldsymbol{x}) \cdot \nabla\phi_j(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \\ &= \int_{\operatorname{supp}\phi_i \cap \operatorname{supp}\phi_j} a(\boldsymbol{x})\nabla\phi_i(\boldsymbol{x}) \cdot \nabla\phi_j(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}.\end{aligned}$$

In particular: Galerkin matrix $\boldsymbol{A}$ is sparse.

Finite Element Approximation

Finite element assembly

Common procedure in assembling the Galerkin system:

(1) Ignore boundary condition initially, i.e., consider all of V^h with nodal basis

$$\{\phi_1, \phi_2, \dots, \phi_n, \phi_{n+1}, \dots, \phi_{\tilde{n}}\},$$

$\tilde{n} - n$ the number of vertices on the boundary ∂D.
Yields matrix $\tilde{\boldsymbol{A}} \in \mathbb{R}^{\tilde{n} \times \tilde{n}}$, vector $\tilde{\boldsymbol{b}} \in \mathbb{R}^{\tilde{n}}$.

(2) Then eliminate the DOF associated with boundary vertices.
Yields matrix $\boldsymbol{A}$, vector $\boldsymbol{b}$.

Note:

- Initial approach for step (1): compute $\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}}$, entry by entry, i.e., basis function by basis function
- But: shape and connectivity of supports typically very different.
- Simpler: compute $\boldsymbol{A}$, $\boldsymbol{b}$ element by element.

Finite Element Approximation

Finite element assembly

$K \in \mathscr{T}^h$: then for $i, j = 1, 2 \ldots, \tilde{n}$:

$$a(\phi_j, \phi_i) = \int_D a \nabla \phi_j \cdot \nabla \phi_i \, \mathrm{d}\boldsymbol{x} = \sum_{K \in \mathscr{T}^h} \int_K a \nabla \phi_j \cdot \nabla \phi_i \, \mathrm{d}\boldsymbol{x} =: \sum_{K \in \mathscr{T}^h} a_K(\phi_j, \phi_i),$$

$$\ell(\phi_i) = \int_D f \phi_i \, \mathrm{d}\boldsymbol{x} = \sum_{K \in \mathscr{T}^h} \int_K f \phi_i \, \mathrm{d}\boldsymbol{x} =: \sum_{K \in \mathscr{T}^h} \ell_K(\phi_i).$$

Setting

$$\begin{aligned} [\tilde{\boldsymbol{A}}_K]_{i,j} &:= a_K(\phi_j, \phi_i) & i, j &= 1, 2, \ldots, \tilde{n}, \\ [\tilde{\boldsymbol{b}}_K]_i &:= \ell_K(\phi_i, & i &= 1, 2, \ldots, \tilde{n}, \end{aligned}$$

we obtain

$$\tilde{\boldsymbol{A}} = \sum_{K \in \mathscr{T}^h} \tilde{\boldsymbol{A}}_K, \qquad \tilde{\boldsymbol{b}} = \sum_{K \in \mathscr{T}^h} \tilde{\boldsymbol{b}}_K.$$

Finite Element Approximation

Finite element assembly: element table

Since each element belongs to the support of exactly three basis functions, only (at most) nine entries of $\tilde{\boldsymbol{A}}_K$ and three entries of $\tilde{\boldsymbol{b}}_K$ are nonzero.
Which entries these are can be determined by maintaining an element table:

$$[ET(i,j)]_{i=1,2,3;j=1,\dots,n_K}:$$

Element	K_1	K_2	$\dots$	K_{n_K}
first vertex	$i_1^{(1)}$	$i_1^{(2)}$	$\dots$	$i_1^{(n_K)}$
second vertex	$i_2^{(1)}$	$i_2^{(2)}$	$\dots$	$i_2^{(n_K)}$
third vertex	$i_3^{(1)}$	$i_3^{(2)}$	$\dots$	$i_3^{(n_K)}$

Here n_K denotes the number of triangles in $\mathscr{T}^h$.

Besides the global vertex numbering

$$x_1, x_2, \dots, x_{\tilde{n}},$$

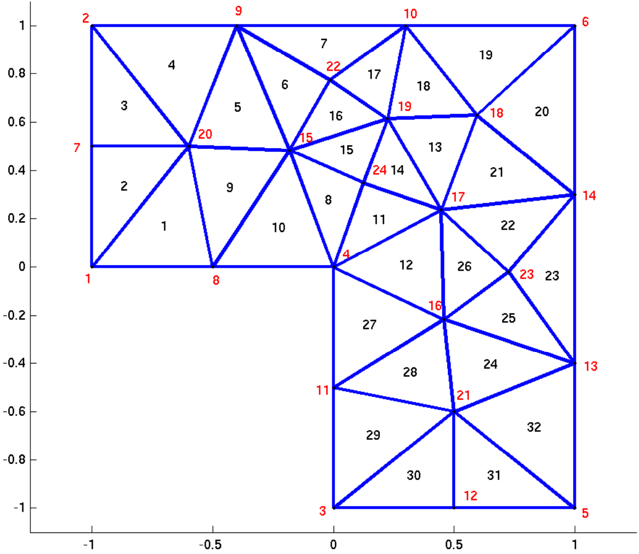
the element table introduces a second, local vertex numbering

$$x_1^{(K)},\ x_2^{(K)},\ x_3^{(K)}$$

of the vertices (DOFs) associated with K.

Finite Element Approximation

Finite element assembly



Global numbering of vertices (red) and elements (black) in a triangulation of an L-shaped domain.

With this notation the nonzero submatrix $\boldsymbol{A}_K$ of $\tilde{\boldsymbol{A}}_K$ and nonzero subvector $\boldsymbol{b}_K$ of $\tilde{\boldsymbol{b}}_K$ are given by

$$\boldsymbol{A}_K := \begin{bmatrix} a_K(\phi_1^{(K)}, \phi_1^{(K)}) & a_K(\phi_2^{(K)}, \phi_1^{(K)}) & a_K(\phi_3^{(K)}, \phi_1^{(K)}) \\ a_K(\phi_1^{(K)}, \phi_2^{(K)}) & a_K(\phi_2^{(K)}, \phi_2^{(K)}) & a_K(\phi_3^{(K)}, \phi_2^{(K)}) \\ a_K(\phi_1^{(K)}, \phi_3^{(K)}) & a_K(\phi_2^{(K)}, \phi_3^{(K)}) & a_K(\phi_3^{(K)}, \phi_3^{(K)}) \end{bmatrix}, \quad \boldsymbol{b}_K := \begin{bmatrix} \ell_K(\phi_1^{(K)}) \\ \ell_K(\phi_2^{(K)}) \\ \ell_K(\phi_3^{(K)}) \end{bmatrix}.$$

If K has number k in the enumeration of the elements, then the association of the local numbering $\{\phi_i^{(K)}\}_{i=1,2,3}$ of the three basis functions whose support contains K with the global numbering $\{\phi_j\}_{j=1}^{\tilde{n}}$ of all basis functions is given by

$$\phi_i^{(K)} = \phi_j, \qquad j = ET(i,k), \quad i = 1,2,3.$$

$\boldsymbol{A}_K$ and $\boldsymbol{b}_K$ are sometimes called the element stiffness matrix and element load vector.

Finite Element Approximation

Finite element assembly

We summarize phase (1) of the finite element assembly process in the following algorithm[4]

Algorithm 2: Phase (1) of finite element assembly.

```
Initialize $\tilde{\boldsymbol{A}} := \boldsymbol{O}$, $\tilde{\boldsymbol{b}} := \boldsymbol{0}$.
foreach $K \in \mathscr{T}_h$ do
    Compute $\boldsymbol{A}_K$ and $\boldsymbol{b}_K$
    $k \leftarrow$ [index of element $K$]
    $i_1 \leftarrow ET(1,k)$, $i_2 \leftarrow ET(2,k)$, $i_3 \leftarrow ET(3,k)$
    $\tilde{\boldsymbol{A}}([i_1 i_2 i_3],[i_1 i_2 i_3]) \leftarrow \tilde{\boldsymbol{A}}([i_1 i_2 i_3],[i_1 i_2 i_3]) + \boldsymbol{A}_K$
    $\tilde{\boldsymbol{b}}([i_1 i_2 i_3]) \leftarrow \tilde{\boldsymbol{b}}([i_1 i_2 i_3]) + \boldsymbol{b}_K$
```

[4] We use the following Matlab-inspired notation:

$$\boldsymbol{A}([i_1 i_2 i_3],[i_1 i_2 i_3]) = \begin{bmatrix} a_{i_1,i_1} & a_{i_1,i_2} & a_{i_1,i_3} \\ a_{i_2,i_1} & a_{i_2,i_2} & a_{i_2,i_3} \\ a_{i_3,i_1} & a_{i_3,i_2} & a_{i_3,i_3} \end{bmatrix}, \quad \boldsymbol{b}([i_1 i_2 i_3]) = \begin{bmatrix} b_{i_1} \\ b_{i_2} \\ b_{i_3} \end{bmatrix}.$$

Finite Element Approximation

Reference element

Both the numerical integration as well as the error analysis benefit from a change of variables to a reference element $\hat{K} \subset \mathbb{R}^2$. Each element $K \in \mathscr{T}^h$ then has a parametrization $K = F_K(\hat{K})$, where

$$F_K : \hat{K} \to K, \qquad \hat{K} \ni \boldsymbol{\xi} \mapsto \boldsymbol{x} \in K, \quad \boldsymbol{x} = F_K(\boldsymbol{\xi}) = B_K \boldsymbol{\xi} + \boldsymbol{b}_K.$$

Most common for triangular elements: unit simplex

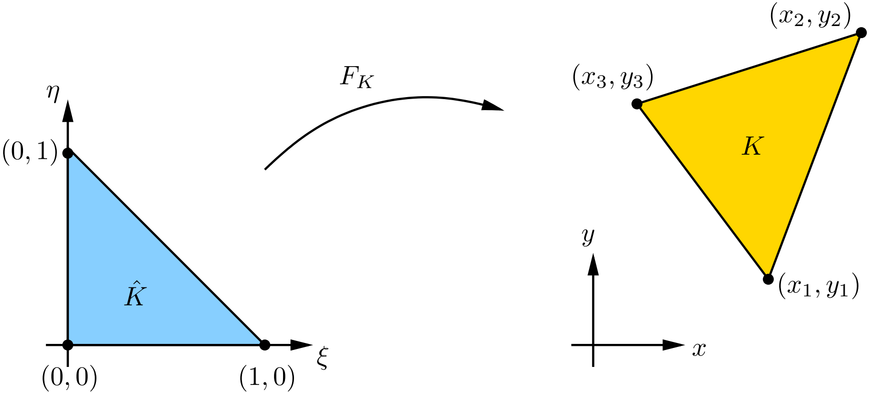
$$\hat{K} = \{(\xi, \eta) \in \mathbb{R}^2 : 0 \leqslant \xi \leqslant 1, 0 \leqslant \eta \leqslant 1 - \xi\}.$$

For each triangle $K \in \mathscr{T}^h$ the affine mapping F_K is determined by prescribing, e.g.,

$$\begin{aligned} (1,0) &\mapsto (x_1, y_1), \\ (0,1) &\mapsto (x_2, y_2), \\ (0,0) &\mapsto (x_3, y_3), \quad \text{i.e.} \end{aligned}$$

Finite Element Approximation

Reference element



$$\begin{bmatrix} x \\ y \end{bmatrix} = \underbrace{\begin{bmatrix} x_1 - x_3 & x_2 - x_3 \\ y_1 - y_3 & y_2 - y_3 \end{bmatrix}}_{B_K} \begin{bmatrix} \xi \\ \eta \end{bmatrix} + \underbrace{\begin{bmatrix} x_3 \\ y_3 \end{bmatrix}}_{\boldsymbol{b}_K}$$

Finite Element Approximation

Reference element

Local (nodal) basis on $\hat{K}$: (dual basis of DOF)

$$\hat\phi_1(\xi,\eta) = \xi, \quad \hat\phi_2(\xi,\eta) = \eta, \quad \hat\phi_3(\xi,\eta) = 1-\xi-\eta, \qquad (\xi,\eta) \in \hat{K}.$$

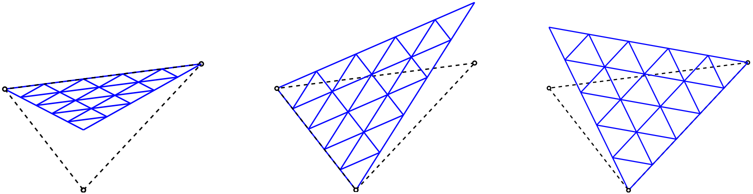
The correspondence

$$\hat\phi \mapsto \phi := \hat\phi \circ F_K^{-1}, \quad \text{d.h.} \quad \phi(\boldsymbol{x}) := \hat\phi(\boldsymbol{\xi}(\boldsymbol{x})) = \hat\phi(F_K^{-1}(\boldsymbol{x}))$$

assigns to $\hat\phi$ on $\hat{K}$ a unique function ϕ on K.

Local basis functions on K:

$$\phi_j = \hat\phi_j \circ F_K^{-1} : K \to \mathbb{R}, \qquad j = 1,2,3.$$



Finite Element Approximation

Reference element, change of variables

The chain rule[5] applied to $\phi(\boldsymbol{x}) = \hat{\phi}(\boldsymbol{\xi}(\boldsymbol{x}))$ gives

$$\nabla\phi = \begin{bmatrix} \phi_x \\ \phi_y \end{bmatrix} = \begin{bmatrix} \hat{\phi}_\xi \xi_x + \hat{\phi}_\eta \eta_x \\ \hat{\phi}_\xi \xi_y + \hat{\phi}_\eta \eta_y \end{bmatrix} = \begin{bmatrix} \xi_x & \eta_x \\ \xi_y & \eta_y \end{bmatrix} \begin{bmatrix} \hat{\phi}_\xi \\ \hat{\phi}_\eta \end{bmatrix} = (DF_K^{-1})^\top \hat{\nabla}\hat{\phi}.$$

Since
$$\begin{aligned} \boldsymbol{x} &= F_K(\boldsymbol{\xi}) = B_K\boldsymbol{\xi} + \boldsymbol{b}_K, &&\text{i.e. } DF_K \equiv B_K, \\ \boldsymbol{\xi} &= F_K^{-1}(\boldsymbol{x}) = B_K^{-1}(\boldsymbol{x} - \boldsymbol{b}_K), &&\text{i.e. } DF_K^{-1} \equiv B_K^{-1} \end{aligned}$$

we obtain

$$\nabla\phi = B_K^{-\top}\hat{\nabla}\hat{\phi}.$$

[5] $\hat{\nabla}$ indicates differentiation with respect to the variables ξ and η.

Finite Element Approximation

Reference element, element integrals

This finally gives the element integrals ($\phi_i = \phi_i^{(K)}$, $i = 1, 2, 3$)

$$\begin{aligned} a_K(\phi_j, \phi_i) &= \int_K a(\boldsymbol{x})\, \nabla\phi_j(\boldsymbol{x}) \cdot \nabla\phi_i(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\ &= \int_{\hat{K}} a(\boldsymbol{x}(\boldsymbol{\xi}))\ \left(B_K^{-\top} \hat{\nabla}\hat{\phi}_j(\boldsymbol{\xi})\right) \cdot \left(B_K^{-\top} \hat{\nabla}\hat{\phi}_i(\boldsymbol{\xi})\right) |\det B_K|\, \mathrm{d}\boldsymbol{\xi}. \end{aligned} \tag{B.11}$$

The determinant is given by (note K is a triangle)

$$|\det B_K| = 2|K|,$$

$$B_K^{-\top} = \frac{1}{2|K|} \begin{bmatrix} y_2 - y_3 & x_3 - x_2 \\ y_3 - y_1 & x_1 - x_3 \end{bmatrix},$$

$$\begin{bmatrix} \hat{\nabla}\hat{\phi}_1 & \hat{\nabla}\hat{\phi}_2 & \hat{\nabla}\hat{\phi}_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}.$$

Finite Element Approximation

Eliminate constrained boundary DOF

To impose the Dirichlet boundary condition we require that the Galerkin approximation $u^h \in V^h$ satisfy

$$u^h(\boldsymbol{x}_j) = g(\boldsymbol{x}_j) \quad \text{at all boundary vertices } \{\boldsymbol{x}_j\}_{j=n+1}^{\tilde{n}}. \tag{B.12}$$

- We partition the coefficient vector $\boldsymbol{u} \in \mathbb{R}^{\tilde{n}}$ into a first block $\boldsymbol{u}_I \in \mathbb{R}^n$ containing the coefficients associated with the interior vertices $\{\boldsymbol{x}_j\}_{j=1}^{n}$ and a second block $\boldsymbol{u}_B \in \mathbb{R}^{\tilde{n}-n}$ containing the constrained coefficients associated with boundary vertices.
- For the assembled matrix $\tilde{\boldsymbol{A}}$ and vector $\tilde{\boldsymbol{b}}$ this induces the partitionings

$$\tilde{\boldsymbol{A}} = \begin{bmatrix} \tilde{\boldsymbol{A}}_{II} & \tilde{\boldsymbol{A}}_{IB} \\ \tilde{\boldsymbol{A}}_{BI} & \tilde{\boldsymbol{A}}_{BB} \end{bmatrix}, \qquad \tilde{\boldsymbol{b}} = \begin{bmatrix} \tilde{\boldsymbol{b}}_I \\ \tilde{\boldsymbol{b}}_B \end{bmatrix}.$$

- The constraint (B.12) now reads $\boldsymbol{u}_B = \boldsymbol{g}$, where $\boldsymbol{g} \in \mathbb{R}^{\tilde{n}-n}$ contains the boundary data $\{g(\boldsymbol{x}_j)\}_{j=n+1}^{\tilde{n}}$.

Finite Element Approximation

Eliminate constrained boundary DOF

This constraint is characterized by there being no coupling of the boundary DOF to either interior DOF or among themselves, resulting in the modified linear system of equations

$$\begin{bmatrix} \tilde{\boldsymbol{A}}_{II} & \tilde{\boldsymbol{A}}_{IB} \\ \boldsymbol{O} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_I \\ \boldsymbol{u}_B \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_I \\ \boldsymbol{g} \end{bmatrix},$$

which gives the reduced system

$$\boldsymbol{A}\boldsymbol{u}_I = \boldsymbol{b}, \qquad \boldsymbol{A} = \tilde{\boldsymbol{A}}_{II}, \quad \boldsymbol{b} = \boldsymbol{b}_I - \tilde{\boldsymbol{A}}_{IB}\boldsymbol{g}$$

for the interior DOF.

Note that this procedure is a discrete variant of the reformulation of the BVP with inhomogeneous Dirichlet boundary conditions to an equivalent one with homogeneous Dirichlet boundary conditions.

Contents

5 Probability Theory

6 Elliptic Boundary Value Problems

6.1 Weak Formulation

6.2 Finite Element Approximation

6.3 Finite Element Convergence

7 Collection of Results from Functional Analysis

8 Miscellanea

Finite Element Convergence

...in a nutshell

- Céa's lemma characterizes the Galerkin error as one of best appproximation from the FE subspace V^h.
- An upper bound for this error is the distance of the true solution from its interpolant from the FE subspace. This is the uniquely determined function from V^h which possesses the same global DOF as the exact solution.
- The asymptotic behavior of the interpolant is then analyzed on a sequence of meshes $\{\mathscr{T}_{h_n}\}_{n\in\mathbb{N}}$ with $\lim_{n\to\infty} h_n = 0$.
- For the interpolation error to become small, the mesh sequence has to be shape-regular: if ρ_K denotes the radius of the inscribed circle in K and $h_K = \operatorname{diam} K$, then a sequence of meshes is shape-regular provided the ratio

  $$\frac{\rho_K}{h_K}, \qquad K \in \mathscr{T}_h$$

  is bounded below uniformly for all $\{\mathscr{T}_{h_n}\}$.
- A priori convergence bounds are obtained by relating the smoothness of the exact solution to the convergence rate h^α of the interpolation error as $h \to 0$.

Finite Element Convergence

Extra regularity

Interpolation estimates for a solution u which is only in $H^1(D)$ do not yield a useful rate h^α with an $\alpha > 0$. For this reason one usually tries to show that the solution possesses more regularity.

Definition B.16

For $r \in \mathbb{N}$ and $D \subset \mathbb{R}^d$ bounded, we denote by $H^r(D)$ the Sobolev space

$$H^r(D) := \{v \in L^2(D) : D^{\boldsymbol{\alpha}} u \in L^2(D) \text{ for all } \boldsymbol{\alpha} \in \mathbb{N}_0^d, |\boldsymbol{\alpha}| \leqslant r\}$$

$H^r(D)$ is a Hilbert space with the inner product

$$(u, v)_{H^r(D)} = \sum_{|\boldsymbol{\alpha}| \leqslant r} \int_D (D^{\boldsymbol{\alpha}} u)(D^{\boldsymbol{\alpha}} v) \, \mathrm{d}\boldsymbol{x}.$$

Finite Element Convergence

Extra regularity, fractional index

For any $r \in \mathbb{R}\backslash\mathbb{N}_0$ we set $r = k + s$, $k \in \mathbb{N}_0$, $s \in (0,1)$ and denote by $|\cdot|_{H^r(D)}$ and $\|\cdot\|_{H^r(D)}$ the Sobolev-Slobodetskii semi-norm and norm defined for $v \in H^k(D)$ by

$$|v|_{H^r(D)} = \left(\int_{D\times D} \sum_{|\boldsymbol{\alpha}|=k} \frac{[D^{\boldsymbol{\alpha}} v(\boldsymbol{x}) - D^{\boldsymbol{\alpha}} v(\boldsymbol{y})]^2}{|\boldsymbol{x}-\boldsymbol{y}|^{d+2s}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \right)^{1/2} \quad \text{and}$$

$$\|v\|_{H^r(D)} = \left(\|v\|_{H^k(D)}^2 + |v|_{H^r(D)}^2 \right)^{1/2}.$$

The Sobolev space $H^r(D)$ is then defined as the space of functions $v \in H^k(D)$ such that $|v|_{H^r(D)}^2$ is finite.

Finite Element Convergence

Interpolation error of linear FE for H^2-regular functions

- Let V^h denote the space of piecewise linear functions subject to a shape-regular, admissible triangulation $\mathscr{T}_h$ of D.
- Denote by $I_h : C(\overline{D}) \to V^h$ the (global) interpolation operator assigning to each continuous function v the interpolant $v_h \in V^h$ determined by the condition that v_h agrees with v at all vertices of $\mathscr{T}_h$.
- Then the error of best approximation of $u \in C(\overline{D})$ is bounded by the interpolation error

$$\inf_{v \in V^h} |u - v|_{H^1(D)} \leqslant |u - I_h u|_{H^1(D)}.$$

- If the solution u of (B.4) has additional regularity $u \in H^2(D)$, then the Sobolev imbedding theorem assures that u agrees a.e. with a function in $C(\overline{D})$, so that pointwise evaluation of u and thus the interpolant is well-defined.
- In this case a scaling argument can be used to show

$$|u - I_h u|_{H^1(D)} \leqslant K\, h\, |u|_{H^2(D)}$$

with a constant K independent of h and u.

Assumption B.17 (H^2 regularity)

There exists a constant $K_2 > 0$ such that, for every $f \in L^2(D)$, the solution of (B.4) belongs to $H^2(D)$ and satisfies

$$|u|_{H^2(D)} \leqslant K_2 \|f\|_{L^2(D)}.$$

Theorem B.18

Under Assumptions B.3 and B.17, the solution u of (B.4) with $f \in L^2(D)$ and the piecewise linear finite element approximation u_h on a sequence of shape-regular meshes satisfy

$$|u - u_h|_a \leqslant K\sqrt{a_{\max}}|u|_{H^2(D)}\, h \leqslant KK_2\sqrt{a_{\max}}\|f\|_{L^2(D)}\, h \tag{B.13}$$

with a constant K independent of h.

Corollary B.19

Under the assumptions of Theorem B.18 there holds

$$|u - u_h|_{H^1(D)} \leqslant K\sqrt{\frac{a_{\max}}{a_{\min}}}|u|_{H^2(D)}\, h \leqslant KK_2\sqrt{\frac{a_{\max}}{a_{\min}}}\|f\|_{L^2(D)}\, h.$$

Finite Element Convergence

Model problem, approximate data

When the coefficient function a and the source term f are replaced by approximations $\tilde{a} \approx a$ and $\tilde{f} \approx f$, then with the modified bilinear and linear forms defined as in (B.6), we may consider the discrete problem

$$\tilde{a}(\tilde{u}_h, v) = \tilde{\ell}(v) \quad \forall v \in V^h. \tag{B.14}$$

In analogy to Theorem B.11 we obtain

Theorem B.20

Under Assumption B.3 let $\tilde{f} \in L^2(D)$ and $g \in H^{1/2}(\partial D)$. Then (B.14) has a unique solutiuon $\tilde{u}_h \in V^h$.

By the triangle inequality, we have

$$|u - \tilde{u}_h|_{H^1(D)} \leqslant |u - \tilde{u}|_{H^1(D)} + |\tilde{u} - \tilde{u}_h|_{H^1(D)}.$$

By an obvious extension of Corollary B.14, we obtain the bound

$$|\tilde{u} - \tilde{u}_h|_{H^1(D)} \leqslant \sqrt{\frac{\tilde{a}_{\max}}{\tilde{a}_{\min}}} \inf_{v \in V^h} |\tilde{u} - v|_{H^1(D)}.$$

Finite Element Convergence

Model problem, approximate data

Alternatively, if we approximate the data at the discrete level only, we may consider the following splitting as more natural:

$$|u - \tilde{u}_h|_{H^1(D)} \leqslant |u - u_h|_{H^1(D)} + |u_h - \tilde{u}_h|_{H^1(D)}.$$

The second term arises, e.g., if we approximate the Galerkin approximation u_h by approximating the bilinear and linear forms using, e.g., piecewise constant approximations of the coefficient a and source term f.

Straightforward modification of the proof of Theorem B.12 yields

$$|u - \tilde{u}_h|_{H^1(D)} \leqslant C_D \tilde{a}_{\min}^{-1} \|f - \tilde{f}\|_{L^2(D)} + \tilde{a}_{\min}^{-1} \|a - \tilde{a}\|_{L^\infty(D)} |u_h|_{H^1(D)}.$$

Contents

5 Probability Theory

6 Elliptic Boundary Value Problems

7 Collection of Results from Functional Analysis

8 Miscellanea

Contents

5 Probability Theory

6 Elliptic Boundary Value Problems

7 Collection of Results from Functional Analysis

7.1 Hilbert-Schmidt Operators

8 Miscellanea

Hilbert-Schmidt Operators

For normed linear spaces X and Y, we denote by $\mathscr{L}(X,Y)$ the set of all bounded linear operators from X to Y.

Definition C.1

Let X and Y be separable Hilbert spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ and let $\{x_j\}_{n\in\mathbb{N}}$ denote a CONS of X. A linear operator $L : X \to Y$ for which

$$\|L\|_{\mathrm{HS}(X,Y)} := \left(\sum_{j=1}^{\infty} \|Lx_j\|_Y^2\right)^{1/2} < \infty$$

is called a Hilbert-Schmidt operator. We shall write $\|L\|_{\mathrm{HS}}$ if $X = Y$.

Proposition C.2

The mapping $\|\cdot\|_{\mathrm{HS}(X,Y)}$ is a norm, called the Hilbert-Schmidt norm, on the space of all Hilbert-Schmidt operators from X to Y, which we denote by $\mathrm{HS}(X,Y)$. In addition, $(\mathrm{HS}(X,Y), \|\cdot\|_{\mathrm{HS}(X,Y)})$ is Banach space.

Hilbert-Schmidt Operators

Examples

Example C.3

For $X = Y = \mathbb{R}^n$ with the Euclidean norm $\|\cdot\|$, the Hilbert-Schmidt norm of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ coincides with the Frobenius-norm $\|\boldsymbol{A}\|_F^2 = \sum_{i,j=1}^n a_{i,j}^2$.

Example C.4

Define $L \in \mathscr{L}(L^2(0,1))$ by

$$(Lu)(x) = \int_0^x u(y)\,\mathrm{d}y, \qquad u \in L^2(0,1), \qquad x \in (0,1).$$

For the CONS $\{f_j(x) = \sqrt{2}\sin(j\pi x) : j \in \mathbb{N}\}$, we have

$$(Lf_j)(x) = \frac{\sqrt{2}}{j\pi}(1 - \cos(j\pi x)).$$

L is a Hilbert-Schmidt operator since $\|Lf_j\|_{L^2(0,1)} \leqslant \frac{2\sqrt{2}}{j\pi}$.

Hilbert-Schmidt Operators

Integral operators

Lemma C.5

Let H be a separable Hilbert space. If $L \in \mathrm{HS}(H)$, then $\|L\|_{\mathscr{L}(H)} \leqslant \|L\|_{\mathrm{HS}}$. In particular, Hilbert-Schmidt operators are bounded.

Definition C.6

For a domain $D \subset \mathbb{R}^d$ and $k \in L^2(D \times D)$, the integral operator with kernel function k is defined as the linear operator

$$K : u \mapsto (Ku)(\boldsymbol{x}) := \int_D k(\boldsymbol{x}, \boldsymbol{y}) u(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}, \qquad \boldsymbol{x} \in D. \tag{C.1}$$

Theorem C.7

An integral operator with kernel function $k \in L^2(D \times D)$ is a Hilbert-Schmidt operator on $L^2(D)$. Conversely, any Hilbert-Schmidt operator K on $L^2(D)$ can be written in the form (C.1) with $\|K\|_{\mathrm{HS}} = \|k\|_{L^2(D\times D)}$.

Hilbert-Schmidt Operators

Compact operators

Definition C.8

A set B in a Banach space X is said to be compact if every sequence $u_n \subset B$ has a convergent subsequence u_{n_k} with limit $u \in B$.

Definition C.9

A linear operator $L : X \to Y$, where X and Y are Banach spaces, is said to be compact if the image of any bounded set $B \subset X$ has compact closure in Y, i.e., if $\overline{L(B)}^{\|\cdot\|_Y}$ is a compact set in Y for all bounded $B \subset X$.

Theorem C.10

For $k \in L^2(D \times D)$ the associated integral operator K on $L^2(D)$ with kernel function k is a compact operator.

Hilbert-Schmidt Operators

Selfadjoint operators, eigenvalues

Definition C.11

An operator $L \in \mathscr{L}(H)$ on a Hilbert space H is said to be selfadjoint if

$$(Lu, v) = (u, Lv) \qquad \forall u, v \in H.$$

Proposition C.12

For a domain $D \subset \mathbb{R}^d$, if $k \in L^2(D \times D)$ is symmetric, i.e., $k(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in D$, then the integral operator with kernel function k is selfadjoint with respect to the $L^2(D)$ inner product.

Definition C.13

If $L \in \mathscr{L}(H)$, $\lambda \in \mathbb{C}$ is called an eigenvalue of L if there exists nonzero $\phi \in H$ such that $L\phi = \lambda\phi$. The element ϕ is called an eigenvector or eigenfunction of L.

Hilbert-Schmidt Operators

Spectral theorem

Theorem C.14 (Spectral theorem for selfadjoint compact operators)

Let H be a separable Hilbert space and $K \subset \mathscr{L}(H)$ be selfadjoint and compact. Denote the eigenvalues of K by $\{\lambda_j\}_{j\in\mathbb{N}}$ ordered such that $|\lambda_{j+1}| \leqslant |\lambda_j|$ and denote the associated eigenfunctions by $\{\phi_j\}$. Then

(i) All eigenvalues are real and $\lambda_j \to 0$ as $j \to \infty$.

(ii) The sequence $\{\phi_j\}$ can be chosen as a CONS of the range $K(H)$ of K and,

(iii) for any $u \in H$,

$$Ku = \sum_{j=1}^{\infty} \lambda_k (u, \phi_j)\phi_j. \tag{C.2}$$

Definition C.15

A function $k : D \times D \to \mathbb{R}$ is nonnegative definite if for any set of points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in D$ and numbers $a_1, \dots, a_n \in \mathbb{R}$ there holds

$$\sum_{j,k=1}^{n} a_j a_k k(\boldsymbol{x}_j, \boldsymbol{x}_k) \geqslant 0.$$

A linear operator $L \in \mathscr{L}(H)$ on a Hilbert space H is nonnegative definite if

$$(u, Lu) \geqslant 0 \qquad \forall u \in H$$

and positive definite if

$$(u, Lu) > 0 \qquad \forall u \in H.$$

Hilbert-Schmidt Operators

Nonnegative functions, operators, trace class operators

Lemma C.16

For a domain $D \subset \mathbb{R}^d$ and a nonnegative definite function $k \in C(D \times D)$, the integral operator K on $L^2(D)$ with kernel function k is nonnegative.

Lemma C.17 (Dini)

For a bounded domain D let $f_n \in C(\overline{D})$ be such that $f_n(\boldsymbol{x}) \leqslant f_{n+1}(\boldsymbol{x})$ for $n \in \mathbb{N}$ and $f_n(\boldsymbol{x}) \to f(\boldsymbol{x})$ as $n \to \infty$ for all $\boldsymbol{x} \in \overline{D}$. Then $\|f - f_n\|_\infty \to 0$ as $n \to \infty$.

Definition C.18

Let H be a separable Hilbert space. A nonnegative definite operator $L \in \mathscr{L}(H)$ is said to be of trace class if $\operatorname{trace}(L) < \infty$, where the trace of L is defined as

$$\operatorname{trace}(L) := \sum_{j=1}^{\infty} (L\psi_j, \psi_j)$$

for any CONS $\{\psi_j\}_{j \in \mathbb{N}}$ of H.

Hilbert-Schmidt Operators

Mercer's theorem

Theorem C.19 (Mercer)

For a bounded domain D, let $k \in C(\overline{D} \times \overline{D})$ be a symmetric and nonnegative definite function and let K be the integral operator with kernel function k. There exist eigenfunctions ϕ_j of K with eigenvalues $\lambda_j > 0$ such that $\phi_j \in C(\overline{D})$ and

$$k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{\infty} \lambda_j \phi_j(\boldsymbol{x}) \phi_j(\boldsymbol{y}), \qquad \boldsymbol{x}, \boldsymbol{y} \in D,$$

where the series converges in $C(\overline{D} \times \overline{D})$. Furthermore,

$$\sup_{\boldsymbol{x}, \boldsymbol{y} \in \overline{D}} \left| k(\boldsymbol{x}, \boldsymbol{y}) - \sum_{j=1}^{n} \lambda_j \phi_j(\boldsymbol{x}) \phi_j(\boldsymbol{y}) \right| \leqslant \sup_{\boldsymbol{x} \in \overline{D}} \sum_{j=n+1}^{\infty} \lambda_j |\phi_j(\boldsymbol{x})|^2. \tag{C.3}$$

The operator K is of trace class and

$$\operatorname{trace}(K) = \int_D k(\boldsymbol{x}, \boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

Contents

5. Probability Theory

6. Elliptic Boundary Value Problems

7. Collection of Results from Functional Analysis

8. Miscellanea

Contents

Bessel Functions

Bessel functions arise, e.g., when constructing eigenfunction expansions for the Laplacian in cylindrical coordinates. They are solutions to Bessel's differential equation with parameter $\nu \in \mathbb{C}$

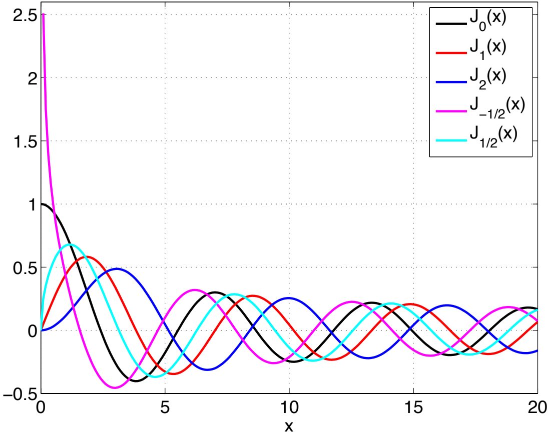
$$u''(z) + \frac{1}{z}u'(z) + \left(1 - \frac{\nu^2}{z_2}\right)u(z) = 0$$

for various boundary conditions.

We are interested in the Bessel functions of the first kind, denoted by $J_\nu(z)$ for real nonnegative argument $z = x \geqslant 0$. These are finite at $x = 0$ for ν nonnegative or integer and singular there for ν negative or non-integer.

For special values of $\boldsymbol{u}$ they possess simple expressions, e.g.

$$J_{-1/2}(x) = \left(\frac{2}{\pi x}\right)^{1/2} \cos x, \qquad J_{1/2}(x) = \left(\frac{2}{\pi x}\right)^{1/2} \sin x,$$



Bessel functions $J_\nu(x)$ may be evaluated in Matlab with the command `besselj(nu,x)`.

Bessel Functions

Fourier transforms spherically symmetric functions

Functions $u : \mathbb{R}^d \to \mathbb{C}$ which are spherically symmetric w.r.t. the origin, i.e., for which

$$u(\boldsymbol{x}) = u(r), \qquad r = \|\boldsymbol{x}\|_2,$$

are sometimes called isotropic functions.

Theorem D.1

The Fourier transform $\hat{u}$ of an isotropic function $u : \mathbb{R}^d \to \mathbb{C}$ is isotropic. Using $u(r)$ and $\hat{u}(\lambda)$ with $\lambda = \|\boldsymbol{\lambda}\|_2$ to denote $u(\boldsymbol{x})$ and $\hat{u}(\boldsymbol{\lambda})$, there holds

$$\hat{u}(\lambda) = (2\pi)^{-d/2} \int_0^\infty J_\nu(\lambda r)(\lambda r)^{-\nu} u(r) r^{d-1}\, \mathrm{d}r, \qquad \nu = \frac{d}{2} - 1.$$

This special case of the Fourier transform is known as a Hankel transform.

Bessel Functions

Modified Bessel functions

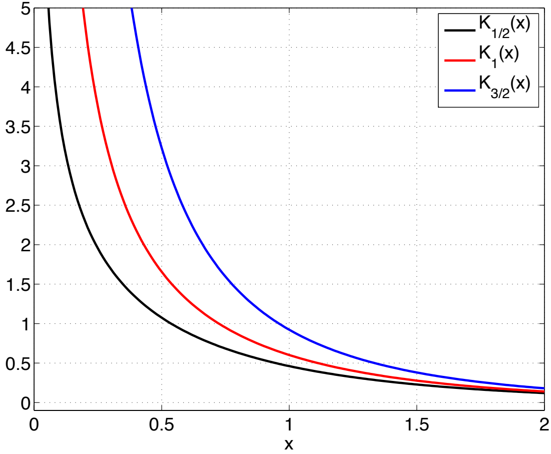
The modified Bessel functions (or hyperbolic Bessel functions) of the first and second kind are given by

$$I_\nu(x) = i^{-\nu} J_\nu(ix), \qquad K_\nu(x) = \frac{\pi}{2} \frac{I_{-\nu}(x) - I_\nu(x)}{\sin(\nu\pi)}.$$

The modified Bessel functions of the second kind K_ν arise in the definition of the Matérn class of isotropic covariance functions.

The special case $\nu = \frac{1}{2}$ has a simple expression:

$$K_{\frac{1}{2}}(x) = \left(\frac{\pi}{2x}\right)^{1/2} e^{-x}, \qquad x \geqslant 0. \tag{D.1}$$



Bessel functions $K_\nu(x)$ may be evaluated in Matlab with the command `besselk(nu,x)`.

Bessel Functions

Modified Bessel functions, half integral order

In view of the following relations which hold for the modified Bessel functions (cf. [Lebedev, 1972])

$$K_{\nu-1}(z) - K_{\nu+1}(z) = -\frac{2\nu}{z} K_\nu(z), \qquad K_{-\nu}(z) = K_\nu(z)$$

along with (D.1), the modified Bessel functions of half-integral order $\nu + \frac{1}{2}$, $\nu \in \mathbb{N}_0$, may be expressed in terms of the exponential function, the square root, and a polynomial in $1/z$. In particular,

$$\begin{aligned} K_{3/2}(z) &= \left(\frac{\pi}{2z}\right)^{1/2} e^{-z} \left(1 + \frac{1}{z}\right), \\ K_{5/2}(z) &= \left(\frac{\pi}{2z}\right)^{1/2} e^{-z} \left(1 + \frac{3}{z} + \frac{3}{z^2}\right). \end{aligned}$$

Bessel Functions

Modified Bessel functions, half integral order, Matérn kernels

For the associated Matérn covariance kernel $c(r) = c(r; \sigma, \nu, \rho)$, we obtain

$$
\begin{aligned}
c(r;\sigma,1/2,\rho) &= \sigma^2 \exp\left(\frac{-\sqrt{2}r}{\rho}\right),\\
c(r;\sigma,3/2,\rho) &= \sigma^2 \exp\left(-\frac{\sqrt{6}r}{\rho}\right)\left(1+\frac{\sqrt{6}r}{\rho}\right),\\
c(r;\sigma,5/2,\rho) &= \sigma^2 \exp\left(-\frac{\sqrt{10}r}{\rho}\right)\left(1+\frac{\sqrt{10}r}{\rho}+\frac{1}{3}\left(\frac{\sqrt{10}r}{\rho}\right)^2\right).
\end{aligned}
$$